Sentiment analysis of Getúlio Vargas' Coup d'état Speech in 1937

Tiago de Almeida Silva

2022-04-22

Introduction

In this project, I will analyze the most common relevant words and the overall sentiment analysis of Getúlio's Vargas famous speech during the Coup d'état in Brazil in 1937. The 1937 Brazilian coup d'état (Portuguese: Golpe de Estado no Brasil em 1937), also known as the Estado Novo coup (Portuguese: Golpe do Estado Novo), was a military coup led by President Getúlio Vargas with the support of the Armed Forces on 10 November 1937. Vargas had risen to power in 1930 with the backing of the military, following a revolution that ended a decades-old oligarchy. Vargas ruled as provisional president until the National Constituent Assembly election in 1934. Under a new constitution, Vargas became the constitutional president of Brazil, but following a 1935 communist uprising, speculation grew over a potential self-coup. Candidates for the 1938 presidential election appeared as early as late 1936. Vargas could not seek re-election, but he and his allies were unwilling to abandon power. Despite loosening political repression after the communist revolt, strong sentiment for a dictatorial government remained, and increasing federal intervention in state governments would pave the way for a coup to take place.

Installing and loading the packages and libraries that will be used in this project:

```r
pacotes <- c("tidytext","ggplot2","dplyr","tibble","gutenbergr","wordcloud",
             "stringr","SnowballC","widyr","janeaustenr","lexiconPT",
             "tidyr","readxl","tm","e1071","gmodels","caret","reshape2", "stringi")

if(sum(as.numeric(!pacotes %in% installed.packages())) != 0){
  instalador <- pacotes[!pacotes %in% installed.packages()]
  for(i in 1:length(instalador)) {
    install.packages(instalador, dependencies = T)
    break()}
  sapply(pacotes, require, character = T)
} else {
  sapply(pacotes, require, character = T)
}
```

Loading and assigning the Vargas' speech to an object named "vargas_speech":

```r
vargas_speech <- read.delim("GetulioVargas_speech.txt", header = F, encoding = "UTF-8") %>%
  rename ("text" = "V1")

vargas_speech %>%
  head(15) %>%
  knitr::kable()
```

text
À nação
O homem de Estado, quando as circunstâncias impõem uma decisão excepcional, de amplas repercussões e profundos efeitos na vida do
país, acima das deliberações ordinárias da atividade governamental, não pode fugir ao dever de tomá-la, assumindo, perante a sua consciência e
a consciência dos seus concidadãos as responsabilidades inerentes à alta função que lhe foi delegada pela confiança nacional.
A investidura na suprema direção dos negócios públicos não envolve, apenas, a obrigação de cuidar e prover as necessidades imediatas e
comuns da administração. As exigências do momento histórico e as solicitações do interesse coletivo reclamam, por vezes, imperiosamente, a
adoção de medidas que afetam os pressupostos e convenções do regime, os próprios quadros institucionais, os processos e métodos de governo.
Por certo, essa situação especialíssima só se caracteriza sob aspectos graves e decisivos nos períodos de profunda perturbação política, econômica
e social. A contingência de tal ordem chegamos, infelizmente, como resultante de acontecimentos conhecidos, estranhos à ação governamental,
que não os provocou nem dispunha de meios adequados para evitá-los ou remover-lhes as funestas consequências.
Oriundo de um movimento revolucionário de amplitude nacional e mantido pelo poder constituinte da nação, o governo continuou, no
período legal, a tarefa encetada de restauração econômica e financeira e, fiel às convenções do regime, procurou criar, pelo alheamento às competições
partidárias, uma atmosfera de serenidade e confiança, propícia ao desenvolvimento das instituições democráticas.
Enquanto assim procedia, na esfera estritamente política, aperfeiçoava a obra de justiça social a que se votara desde o seu advento, pondo
em prática um programa isento de perturbações e capaz de atender às justas reivindicações das classes trabalhadoras, de preferência as concernentes

Part 1 - Creating a word cloud of the most common words in Vargas' Speech

Separating and assigning each word (token) in the speech to a different and unique row:

This method is known as *Tokenization* which is the process of splitting a phrase, sentence, paragraph, or one or multiple text documents into smaller units.

```
speech_token <- vargas_speech %>%
  unnest_tokens(word, text) %>%
  count(word, sort = T)

speech_token %>%
  head(15) %>%
  knitr::kable()
```

word	n
de	181
a	136
e	125
o	94
que	83

word	n
da	67
do	66
se	56
em	51
as	45
os	45
não	39
dos	37
das	36
à	26

Excluding the stop words in Portuguese. That's important to not have any noise in the sentiment analysis as words like articles and prepositions do not give us any sense of sentiment or feelings.

Code which shows the package with some of the stop words in Portuguese that will be used in this project:

```
head(stopwords::stopwords("portuguese"), 40)
```

```
##  [1] "de"      "a"      "o"       "que"    "e"       "do"      "da"       "em"
##  [9] "um"      "para"   "com"     "não"    "uma"     "os"      "no"       "se"
## [17] "na"      "por"    "mais"    "as"     "dos"     "como"    "mas"      "ao"
## [25] "ele"     "das"    "à"       "seu"    "sua"     "ou"      "quando"   "muito"
## [33] "nos"     "já"     "eu"      "também" "só"      "pelo"    "pela"     "até"
```

Removing the stop words from different sources through anti_joins:

```
speech_final <-  speech_token %>%  anti_join(get_stopwords(language = "pt",
                                             source = "snowball"), by = "word")


speech_final <- speech_final %>% anti_join(get_stopwords(language = "pt",
                                           source = "nltk"),by = "word")
speech_final <- speech_final %>% anti_join(get_stopwords(language = "pt",
                                           source = "stopwords-iso"),by = "word")
```

I will create a word cloud to get a graphic overview of the most common relevant words in Vargas' speech:

```
pal <- brewer.pal(8,"Dark2")

speech_final %>%
  with(wordcloud(word, n, random.order = F, max.words = 50, colors = pal))
```

The most common words in his speech are nationalist ones, and that's very understandable because he was trying to get the country together because of his Coup d'état bad image in society. Words like *nacional* / *national*, *nação* / *nation*, *país* / *country* and *vida* / *life* appeared many times in the speech's text, and it shows how Vargas were trying to convince people that what he did was the best for the country.

Part 2 - Getting the Sentiment Analysis of Vargas' Speech

There are dictionaries that show scores for every single word, and they contain information about the emotions or polarity expressed by words, phrases, or concepts. In practice, a dictionary usually provides one or more scores for each word. We can then use them to compute the overall sentiment of an input sentence based on individual words.

Before I start my analysis, it is important to keep in mind that there are some limitations with this sort of analysis (based on words in a dataframe), especially in Portuguese because there are very limited dictionaries available in the language and that turns the analysis process bit harder and less accurate. Getting the right syntax context of each word is not always correct and for this reason, the outcome stating if the text is more

positive or negative can be very subjective or interpretative. That's one of the big issues when dealing with human sentiments based on words.

Getting the sentiment dictionary that will be used in this project:

- lexiconPT Dictionary

```
#loading a portuguese sentiment dictionary named lexiconPT

#datasets

data("oplexicon_v3.0")
data("sentiLex_lem_PT02")

#assigning the datasets to objects

op30 <- oplexicon_v3.0
op30 <-  op30 %>%
  rename("word" = "term")

sent_df <- sentiLex_lem_PT02
sent_df <- sent_df %>%
  rename("word" = "term")
```

Taking a look at the op30 dataset:

```
op30 %>%
  head(15) %>%
  knitr::kable()
```

word	type	polarity	polarity_revision
=[	emot	-1	A
=@	emot	-1	A
=p	emot	-1	A
=P	emot	-1	A
=x	emot	-1	A
=d	emot	1	A
=D	emot	1	A
;)	emot	1	A
;)	emot	1	A
;@	emot	-1	A
;*	emot	-1	A
;**	emot	-1	A
;~	emot	-1	A
;D	emot	-1	A
;D	emot	-1	A

Taking a look at the sent dataset:

```
sent_df  %>%
  head(15) %>%
  knitr::kable()
```

word	grammar_category	polarity	polarity_target	polarity_classification
a-vontade	N	1	N0	MAN
abafado	Adj	-1	N0	JALC
abafante	Adj	-1	N0	MAN
abaixado	Adj	-1	N0	JALC
abalado	Adj	-1	N0	JALC
abalizado	Adj	1	N0	JALC
abalroado	Adj	-1	N0	MAN
abalroar	V	1	N0:N1	MAN
abanar	V	1	N0:N1	MAN
abandalhado	Adj	-1	N0	MAN
abandalhamento	N	-1	N0	MAN
abandonado	Adj	-1	N0	JALC
abandonar	V	-1	N0:N1	MAN
abarcante	Adj	0	N0	MAN
abarrotado	Adj	-1	N0	JALC

The sent_df dataset presents less data (rows) than the op30 one but it contains nouns that are very important to my analysis due to nouns have very specific and powerful meanings.

Full joining the op30 and sent_df datasets from the lexiconPT Dictionary:

```
dict <- full_join(x = op30,
                  y = sent_df,
                  by = c("word"))

dict  %>%
  head(15) %>%
  select(everything(), -c(4, 7:8)) %>%
  knitr::kable()
```

word	type	polarity.x	grammar_category	polarity.y
=[	emot	-1	NA	NA
=@	emot	-1	NA	NA
=p	emot	-1	NA	NA
=P	emot	-1	NA	NA
=x	emot	-1	NA	NA
=d	emot	1	NA	NA
=D	emot	1	NA	NA
;)	emot	1	NA	NA
;)	emot	1	NA	NA
;@	emot	-1	NA	NA
;*	emot	-1	NA	NA
;**	emot	-1	NA	NA
;~	emot	-1	NA	NA
;D	emot	-1	NA	NA
;D	emot	-1	NA	NA

As seen above the object dict is kinda confusing because there are two columns for polarity and sometimes the polarities are different for the same word. Moreover, there are lots of NAs that are common in the full joining.

I will create a new column with the mean of the polarities, this way I can get just one sentiment for each word in the dataset. For example, if a word contains a polarity.x = 1 and a polarity.y = -1 (opposite sentiments), in the new column it will have a polarity = 0 which means is a neutral word. That makes sense if we think the word can have such divergent sentiments and it is hard for the dictionary to be sure if a certain word has a positive or negative meaning. In this way, I will only keep the words that have similar polarities.

There will be cases where the polarity value will be 0.5 or -0.5 and I will infer if the word is positive or negative according to its mathematical sign (+ or -). Words in this range can have either a neutral or positive/negative sentiment. As I will not use neutral words in my project due to their lack of useful meaning for Vargas' speech sentiment analysis, I will round them to be 1 or -1.

```
#Creating the average polarity column with both polarities available for each word

dict_complete <- dict %>%
  mutate(polarity = ((polarity.x + polarity.y)/2))

#Assigning existing polarities (polarity.x or polarity.y) to the Nas on average the polarity column

dict_complete$polarity[is.na(dict_complete$polarity)]<-dict_complete$polarity.x[is.na(dict_complete$pol
dict_complete$polarity[is.na(dict_complete$polarity)]<-dict_complete$polarity.y[is.na(dict_complete$pol

dict_complete  %>%
  head(15) %>%
  select(c(1:3, 5:6, 9)) %>%
  knitr::kable()
```

word	type	polarity.x	grammar_category	polarity.y	polarity
=[	emot	-1	NA	NA	-1
=@	emot	-1	NA	NA	-1
=p	emot	-1	NA	NA	-1
=P	emot	-1	NA	NA	-1
=x	emot	-1	NA	NA	-1
=d	emot	1	NA	NA	1
=D	emot	1	NA	NA	1
;)	emot	1	NA	NA	1
;)	emot	1	NA	NA	1
;@	emot	-1	NA	NA	-1
;*	emot	-1	NA	NA	-1
;**	emot	-1	NA	NA	-1
;~	emot	-1	NA	NA	-1
;D	emot	-1	NA	NA	-1
;D	emot	-1	NA	NA	-1

There are some score errors in the polarity.y column as in the dataframe we can see sentiment words on this column scored as:

```
 cat(unique(dict_complete$polarity.y))
```

```
## NA -1 1 0 7 8 -2 -3
```

I will remove the words scored as 7, 8, -2, and -3 due to they are not in the correct range of analysis and they will generate noise in my polarity average column:

7

Ps. there are only 4 words in the dataframe with the outlier scores.

```r
dict_complete <- dict_complete[!(dict_complete$polarity.y %in% c(7, 8, -2, -3) ), ]

#selcting the columns that will be useful in the project

dict_complete <- dict_complete[ ,c(1:2, 5, 9)]

dict_complete  %>%
  head(15) %>%
  knitr::kable()
```

word	type	grammar_category	polarity
=[	emot	NA	-1
=@	emot	NA	-1
=p	emot	NA	-1
=P	emot	NA	-1
=x	emot	NA	-1
=d	emot	NA	1
=D	emot	NA	1
;)	emot	NA	1
;)	emot	NA	1
;@	emot	NA	-1
;*	emot	NA	-1
;**	emot	NA	-1
;~	emot	NA	-1
;D	emot	NA	-1
;D	emot	NA	-1

Taking the mark accents out of the words for the joining:

```r
speech_noMarks <- stri_trans_general(speech_final$word, "Latin-ASCII")

speech_final$noMarks <- speech_noMarks
```

Joining the dictionary to the object to get an overall sentiment score of the Vargas' speech:

```r
sent_score <- speech_final %>%
  inner_join(dict_complete, by = c("noMarks" = "word"))


sent_score %>%
  head(15) %>%
  knitr::kable()
```

word	n	noMarks	type	grammar_category	polarity
nacional	16	nacional	adj	NA	0.0
política	16	politica	adj	NA	0.0
partidos	8	partidos	adj	NA	-1.0

word	n	noMarks	type	grammar_category	polarity
político	7	politico	adj	NA	0.0
problema	7	problema	NA	N	-1.0
consciência	5	consciencia	NA	N	1.0
econômicas	5	economicas	adj	NA	0.0
econômicas	5	economicas	adj	NA	0.0
meios	5	meios	adj	NA	0.0
social	5	social	adj	NA	0.0
armadas	4	armadas	adj	NA	0.0
constitucional	4	constitucional	adj	NA	0.0
econômico	4	economico	adj	Adj	0.5
econômico	4	economico	adj	Adj	0.5
econômico	4	economico	adj	Adj	0.5

As we see, there are some duplicated rows in the dataset so I will remove them and will keep only one of each:

```
sent_score <- unique(sent_score)
```

Getting the average score of the Vargas' speech through the polarity variable. That will give us the first glimpse of the speech's text.

Points to be aware of:

- 1 = *overall positive text content*
- 0 = *overall neutral text content*
- -1 = *overall negative text content*

Special cases in my project:

- 0.5 = *sentiment ranging from neutral to positive*
- -0.5 = *sentiment ranging from neutral to negative*

As I said before I will round the polarity scores above (0.5 and -0.5) respecting their mathematical sign because neutral words will not be used in my project:

```
#Assigning 1 to the 0.5 score

sent_score$polarity[sent_score$polarity == 0.5] <- 1

#Assigning -1 to the -0.5 score

sent_score$polarity[sent_score$polarity == -0.5] <- -1
```

Getting the speech's score:

```
cat(paste("Getulio Vargas' speech score: ",round(mean(sent_score$polarity), 2)))
```

```
## Getulio Vargas' speech score:  0.04
```

As we can see above, the average score of Vargas' speech was **0.04** and that means he had a slightly neutral/positive speech. If we consider he took part in a Coup d'etat and it was his first public speech after the event, we may assume he was trying to be diplomatic and neutral/positive oriented to calm down the population and the press.

I will use the polarity score to determine if a certain word is negative or positive. The documentation says a word has a positive sentiment when it is equal to 1, and a negative one when it is equal to -1.

Ps. I am only interested in the positive and negative words from now on, that's why I am not taking into consideration the neutral ones (0) for the word cloud.

To clarify once more:

- Word = 1, *positive*
- Word = -1, *negative*

```r
sent_score$sentiment[sent_score$polarity == 1] <- "positive"
sent_score$sentiment[sent_score$polarity == -1] <- "negative"

#Removing the NAs concerning the 0 score from the dataset

sent_score <- sent_score[ ,c(1:2, 6:7)]

sent_score_cloud <-  na.omit(sent_score)

sent_score_cloud %>%
  head(15) %>%
  knitr::kable()
```

	word	n	polarity	sentiment
3	partidos	8	-1	negative
5	problema	7	-1	negative
6	consciência	5	1	positive
13	econômico	4	1	positive
17	justiça	4	1	positive
18	maneira	4	1	positive
21	partidárias	4	1	positive
22	perigo	4	-1	negative
26	confiança	3	1	positive
30	guerra	3	-1	negative
31	liberdade	3	1	positive
34	segurança	3	1	positive
35	sentido	3	-1	negative
36	adequado	2	1	positive
37	ameaça	2	-1	negative

A bar plot to see in a graphic way the overall sentiment contained in the text:

```r
plot <-  sent_score_cloud %>%
  group_by(sentiment) %>%
  summarise(total = n()) %>%
  mutate(percentage = paste(round(prop.table(total) * 100, 2), '%'))
```

```
ggplot(plot, aes(x = reorder(sentiment, desc(total)), y = total, fill = sentiment)) +
  geom_col() +
  labs(title = "Frequency of Positive and Negative Words in Vargas' Speech",
       x = "Sentiment",
       y = "Word Count") +
  geom_text(label = plot$percentage, vjust = -0.5) +
  theme_classic()
```

Frequency of Positive and Negative Words in Vargas' Speech
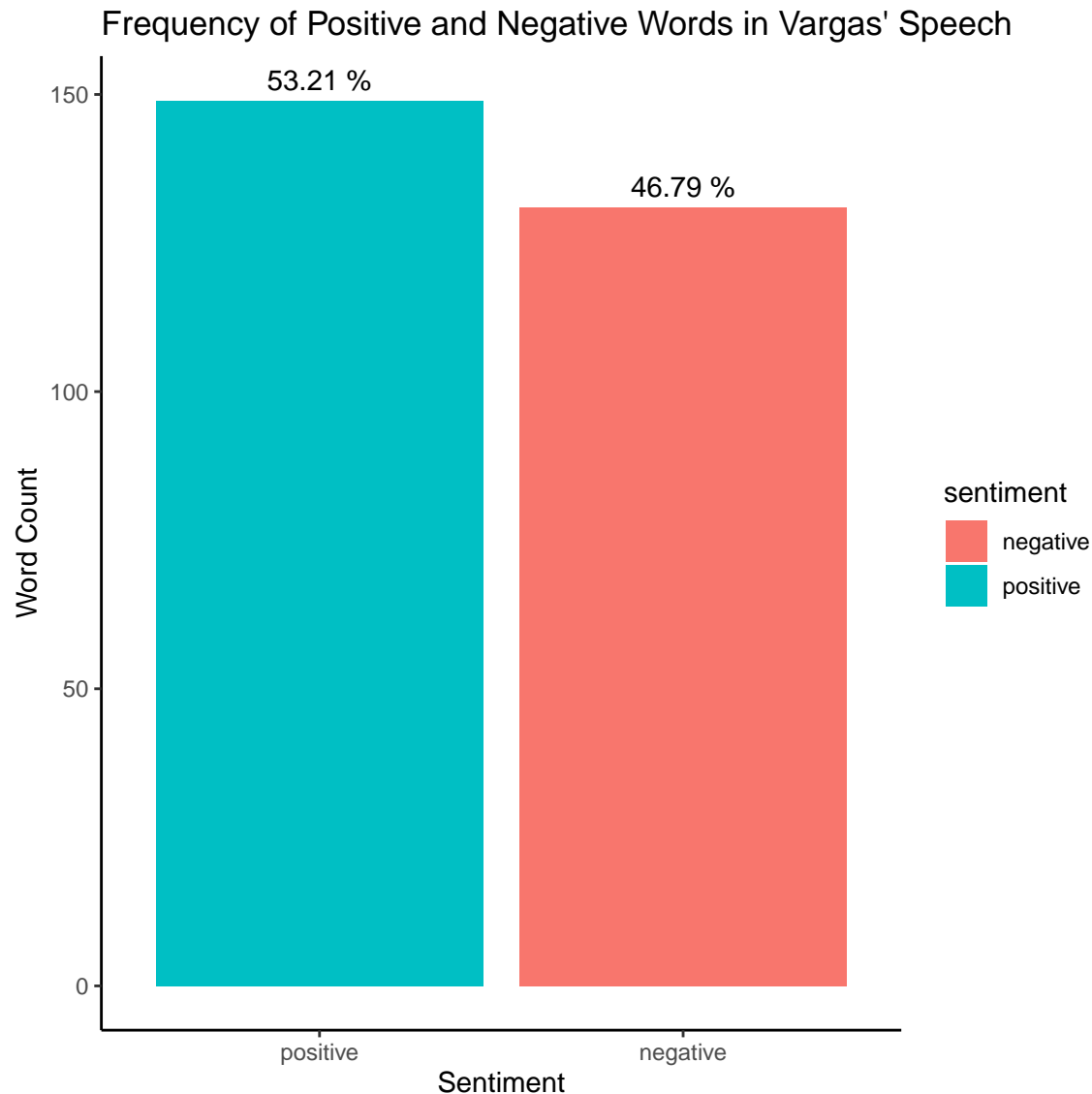
There are more positive words than negative ones in Getulio Vargas' speech but it is interesting to notice the difference between both is not that big.

Creating a word cloud showing the most common positive and negative sentiments found in the text of Getúlio Vargas' speech:

```
sent_score_cloud  %>%
  acast(word ~ sentiment, value.var = "n", fill = 0, fun.aggregate = sum) %>%
  comparison.cloud(colors = c("red2", "darkgreen"),
```

```
            scale = c(6, 0.5),
            max.words = 50,
            title.bg.colors = "white",
            title.size = 3)
```


As seen above it seems there is a slight prevalence of positive words in the top 50 used in the word cloud, but surprisingly the first two most common words are negative ones according to the lexical dataframes. It is necessary to keep in mind, as mentioned before, that there is a limitation with the lexiconPT Dictionary because it only presents some words of the vast Portuguese vocabulary. Moreover, Portuguese is a language with many homographs words (words with different meanings and the same writing) and there is strong evidence that some of the words classified as positive or negative do not belong to their assigned category. I will give one example below taking into consideration the class (adjective, noun, article. . .) of every word contained in the lexiconPT dictionary:

```
dict_complete %>%
  filter(word == "partidos") %>%
  select(everything(), -(3)) %>%
  knitr::kable()
```

word	type	polarity
partidos	adj	-1

- *partidos / parties* (the most common relevant word in the speech), in Varga's speech he uses the word *partidos*, which is an adjective in the dataset, with a meaning of "politic parties" and not "broken" as in Portuguese *partido* (adjective) also has this meaning. In the sentiment analysis, this word was classified as negative but probably it would have a neutral or even a positive sentiment because the word *partidárias / partisans* (derivation of *partido*) is classified as positive in the lexiconPT dataframes.

Another point to be aware of is that the most common words in both word clouds in this project diverge considerably as in the second one only the words in the lexiconPT dataframe with positive or negative sentiments were considered for it. For this reason, the number of available words for part 2 of the project was fewer than in part 1.

Conclusion

Bearing in mind the limitations of the sentiment analysis using lexical dictionaries, the Getúlio Varga's famous speech during his Coup d'etat in 1937 presented a slight prevalence of positive words and that's something very interesting to know about because it is well known that, during the speech day, the environment around him was not peaceful at all and maybe he was trying to get things back on track again. It was also astonishing to see how he used so many nationalist and diplomatic words in his speech text as a clear sign of trying to calm down the population and press. Vargas used to use very tough and straightforward words in his speeches and, in this one, he tried to change the text tone to sound more friendly and centered.