Analisi sui subreddit politici di Reddit

Riccardo Baratto

Contents

Introduzione

Domanda di ricerca

La domanda di ricerca che ha fornito la struttura da seguire per questo progetto è la seguente:

- In che modo i subreddit politicamente polarizzati si differenziano da quelli neutri in termini di temi trattati, complessità linguistica, sentiment e struttura semantica del discorso?

L'analisi verterà quindi sull'esplorazione della comunicazione politica all'interno di sei subreddit con orientamenti politici differenti.

Contesto

I dati sono stati presi da Reddit, che è un sito di social news, intrattenimento e forum, dove degli utenti registrati possono pubblicare dei contenuti.

Reddit ha delle comunità chiamate subreddit per interessi diversi e ogni utente può crearne una.

Gli utenti possono pubblicare immagini o post di testo in linea con le linee guida sui contenuti di Reddit e le regole individuali di ogni subreddit. Possono anche commentare i post di altri.

Mentre gli utenti navigano, possono scegliere di andare a community specifiche o di navigare nella propria homepage, dove trovano i post di tutte le community che seguono.

In questo progetto verrano presi dei post e i commenti sotto esso. Sono stati scelti sei subreddit che trattatano politica USA.

Sono stati scelti i seguenti subreddit di destra:

- Conservative

- Republican

I seguenti subreddit di sinistra:

- Liberal
- democrats

E i seguenti subreddit neutrali:

- PoliticalDiscussion
- politics

Descrizione dati

I dati sono stati raccolti utilizzando la libreria Python PRAW, che permette l'interazioni con l'API ufficiale di Reddit.

Sono stati estratti gli ultimi 1000 post da ciascuno dei sei subreddit selezionati, per un totale di circa 6000 post e fino a 30 commenti per post.

I post sono stati filtrati in modo tale da tenere solo quelli in cui il profilo dell'utente non è stato cancellato, inotre sono poi stati puliti tramite lemmattizzazione usando la libreria "spaCy".

I dati sono stati salvati in due dataset in formato .csv: uno per i post e uno per i commenti.

Metodi utilizzati

Ogni approccio usato è stato scelto per catturare uno specifico aspetto della comunicazione politica nei subreddit.

- Topic modelling
  - Per identificare i temi ricorrenti all'interno dei post, si è scelto di utilizzare il modello di Latent Dirichlet Allocation(LDA). La selezione del numero ottimale di topic è stata guidata da quattro metriche (Griffiths2004, CaoJuan2009, Arun2010, Deveaud2014). Si è scelto di usare il modello LDA poichè consente di estrarre automaticamente i principali argomenti trattati senza supervizione, evidenziando la varietà tematica e la focalizzazione ideologica nei vari subreddit.
- Analisi del Sentiment
  - Il sentiment dei testi è stato misurato usando la libreria "sentimentr", che restituisce un punteggio associato al tono emotivo del post e dei commenti. Si è scelto di misurare il tono generale poichè questo permette di valutare se l'orientamento politico influenzi il registro comunicativo. Le differenze tra gruppi sono state valutate statisticamente tramite il test di Wilcoxon.
- Analisi delle Emozioni
  - Per una rappresentazione più dettagliata del tono emotivo, è stato impiegato il dizionario NRC, che associa le parole a emozioni di base (es. rabbia, paura e tristezza) e atteggiamenti (fiducia e anticipazione). Si è deciso di effettuare anche questa analisi poichè il sentiment score da non solo è sufficiente a cogliere le sfumature emotive.
- Complessità linguistica
  - La complessità dei testi è stata stimata attraverso:
    - Lunghezza media delle frasi
    - Varietà lessicale assoluta(numero di parole uniche)

    - TTR(Type-Token Ratio), ovvero la varietà relativa del vocabolario

  Queste metriche aiutano a valutare il linguaggio usato, infatti una maggiore lunghezza o varietà lessicale può indicare uno stile più argomentativo o riflessivo, mentre frasi brevi e vocabolario ristretto possono suggerire messaggi più diretti o retorici.

- Analisi semantica
  - Per valutare la coerenza e l'organizzazione del discorso:
    - Similarità coseno tra commenti per misurare la ripetitività semantica
    - Coerenza sequenziale dei commenti per stimare la continuità del discorso
    - Entropia lessicale per valutare la complessità informativa

  Si è scelto di utilizzare queste misure poichè permettono di catturare la struttura latente della comunicazione.

- Reti semantiche e co-occorrenza
  - E' stata costruita una rete di co-occorrenza delle parole nei post, calcolando il grado medio dei nodi(frequenza di associazione tra parole) e la densità della rete(interconnesione complessiva dei concetti). La rete semantica consente di quantificare la struttura dei concetti espressi nei post. La densità e il grado medio permettono di comprendere se il discorso è incentrato su pochi concetti ripetuti o se è articolato in una rete più ampia e complessa.

- Parole-ponte tra i gruppi
  - Sono state identificate le parole condivise tra più subreddit per poter individuare elementi lessicali comuni, potenzialmente con significato narrativo trasversale. Le parole-ponte permettono di rilevare i temi su cui gruppi diversi si confrontano, pur da prospettive opposte, e segnalano potenziali spazi di dialogo o conflitto.

Codice

```
knitr::opts_chunk$set(echo = TRUE,
        warning = FALSE, message = FALSE,
        fig.width = 10, fig.height = 6)
```

Librerie

```
library(dplyr)
library(tidytext)
library(ggplot2)
library(topicmodels)
library(tm)
library(SnowballC)
library(textdata)
library(sentimentr)
library(readr)
library(syuzhet)
library(tidyr)
library(tidyverse)
library(text2vec)
library(stringr)
library(data.table)
library(entropy)
library(lubridate)
library(widyr)
library(igraph)
library(ggraph)
library(LDAvis)
library(doParallel)
library(ldatuning)
```

Funzioni

```r
sentiment_analysis <- function(text) {
  sentiment_score <- sentimentr::sentiment(text)$sentiment
  return(mean(sentiment_score, na.rm = TRUE))
}


clean_text <- function(text) {
  text <- tolower(text)
  text <- removePunctuation(text)
  text <- removeNumbers(text)
  text <- removeWords(text, stopwords("en"))
  text <- stripWhitespace(text)
  text <- wordStem(text)
  return(text)
}


find_optimal_topics <- function(dtm, k_min = 2, k_max = 15) {
  result <- FindTopicsNumber(
    dtm,
    topics = seq(k_min, k_max),
    metrics = c("Griffiths2004", "CaoJuan2009",
                "Arun2010", "Deveaud2014"),
    method = "Gibbs",
    control = list(seed = 1123, burnin = 1000, iter = 1000),
    mc.cores = 1L,
    verbose = TRUE
  )

  FindTopicsNumber_plot(result)
  return(result)
}

get_sentence_length <- function(text) {
  words <- unlist(strsplit(text, " "))
  return(length(words))
}


get_vocab_size <- function(text) {
  words <- unlist(strsplit(text, " "))
  return(length(unique(words)))
}
```

```r
get_cosine_similarity <- function(texts) {
  it <- itoken(texts, progressbar = FALSE)
  vectorizer <- vocab_vectorizer(create_vocabulary(it))
  dtm <- create_dtm(it, vectorizer)
  sims <- sim2(dtm, method = "cosine", norm = "l2")
  mean(sims[lower.tri(sims)], na.rm = TRUE)
}

get_thread_coherence <- function(df) {
  df <- df %>% arrange(post_id, created_utc) %>%
    filter(str_count(clean_body, "\\w+") > 1)
  if (nrow(df) < 2) return(NA)
  it <- itoken(df$clean_body, progressbar = FALSE)
  vocab <- create_vocabulary(it)
  if (nrow(vocab) == 0) return(NA)
  vectorizer <- vocab_vectorizer(vocab)
  dtm <- create_dtm(it, vectorizer)
  if (nrow(dtm) < 2) return(NA)
  m <- as.matrix(dtm)
  sims <- sapply(2:nrow(m), function(i) {
    if (sum(m[i,]) == 0 || sum(m[i - 1,]) == 0) return(NA)
    sum(m[i,] * m[i - 1,]) /
      (sqrt(sum(m[i,]^2)) * sqrt(sum(m[i - 1,]^2)))
  })
  mean(sims, na.rm = TRUE)
}

get_entropy <- function(words) {
  freqs <- table(unlist(strsplit(words, " ")))
  entropy(freqs / sum(freqs), unit = "log2")
}


get_network_metrics <- function(edges, min_freq = 10) {
  filtered <- edges %>% filter(n >= min_freq)
  g <- graph_from_data_frame(filtered, directed = FALSE)

  tibble(
    grado_medio = mean(degree(g)),
    densita = edge_density(g),
  )
}
```

Caricamento dei dati

```
posts <- read.csv(file.choose(), sep=";", stringsAsFactors = FALSE)
comments <- read.csv(file.choose(), sep=";", stringsAsFactors = FALSE)
```

Preparazione dei dati

```
#distinzione subreddit in destra, sinistra e neutrale
right_subreddits <- c("Conservative", "Republican")
left_subreddits <- c("Liberal", "democrats")
neutral_subreddits <- c("PoliticalDiscussion", "politics")
right_posts <- posts %>% filter(subreddit %in% right_subreddits)
left_posts <- posts %>% filter(subreddit %in% left_subreddits)
neutral_posts <- posts %>% filter(subreddit %in% neutral_subreddits)
posts_info <- posts %>% select(post_id, subreddit)
comments <- merge(comments, posts_info, by.x = "post_id", by.y = "post_id")
right_comments <- comments %>% filter(subreddit %in% right_subreddits)
left_comments <- comments %>% filter(subreddit %in% left_subreddits)
neutral_comments <- comments %>% filter(subreddit %in% neutral_subreddits)
posts$subreddit_type <- case_when(
  posts$subreddit %in% right_subreddits ~ "Destra",
  posts$subreddit %in% left_subreddits ~ "Sinistra",
  posts$subreddit %in% neutral_subreddits ~ "Neutro"
)
comments <- comments %>%
  mutate(group = case_when(
    subreddit %in% right_subreddits ~ "Destra",
    subreddit %in% left_subreddits ~ "Sinistra",
    subreddit %in% neutral_subreddits ~ "Neutro"
  ))
#associo ogni commento al suo subreddit
posts_info <- posts %>% select(post_id, subreddit)
comments <- comments %>%
  mutate(subreddit_type = case_when(
    subreddit %in% right_subreddits ~ "Destra",
    subreddit %in% left_subreddits ~ "Sinistra",
    subreddit %in% neutral_subreddits ~ "Neutro"
  ))
```

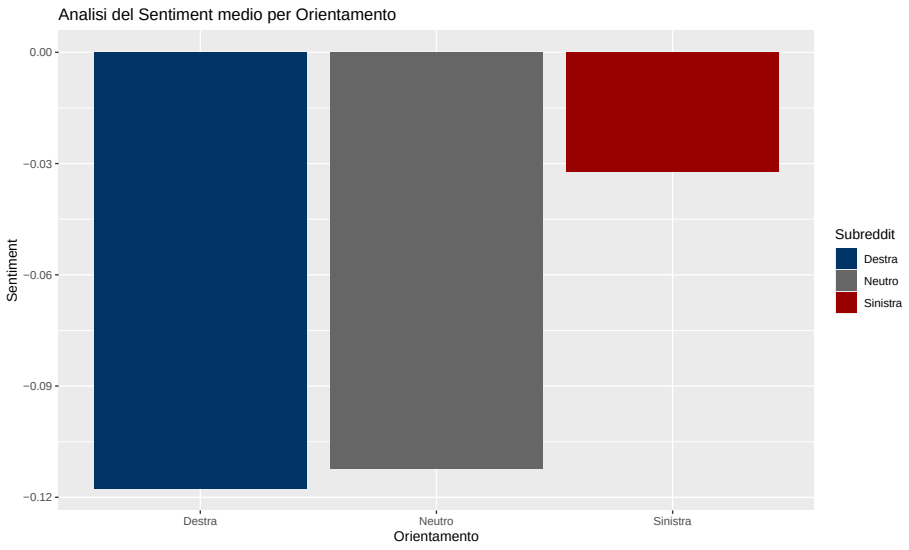
In questa fase si è categorizzato post e commenti distinguendoli tra subreddit orientati a destra, sinistra e neutri. Questa classificazione permetterà di analizzare differenze tematiche, linguistiche ed emotive tra i gruppi.

Sentiment Analysis

```
right_posts$sentiment <- sapply(right_posts$clean_text, sentiment_analysis)
left_posts$sentiment <- sapply(left_posts$clean_text, sentiment_analysis)
neutral_posts$sentiment <- sapply(neutral_posts$clean_text, sentiment_analysis)

#analisi sentiment medio nelle varie tipologie di post
sentiment_summary <- data.frame(
  Subreddit = c("Destra", "Sinistra", "Neutro"),
  Sentiment = c(
    mean(right_posts$sentiment),
    mean(left_posts$sentiment),
    mean(neutral_posts$sentiment)
  )
)
```

Per ciascun post è stato calcolato un punteggio di sentiment, successivamente è stata calcolata la media del sentiment per ciascun gruppo per poterli confrontare.

```
ggplot(sentiment_summary, aes(x=Subreddit, y=Sentiment, fill=Subreddit)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values = c("Destra" = "#003366",
                  "Sinistra" = "#990000",
                  "Neutro" = "#666666")) +
  labs(title="Analisi del Sentiment medio per Orientamento",
       y="Sentiment", x="Orientamento")
```



Tutti i gruppi mostrano un sentiment medio negativo, ma con differenze importanti, in particolare i post dei subreddit di Destra hanno il sentiment medio più negativo, seguiti a poca distanza dai subreddit neutrali. Anche i subreddit di sinistra mostrano un sentiment negativo, ma non a livello dei altri due gruppi. I contenuti dei subreddit di destra sembrebbero essere caratterizzati da un tono più polemico, conflittuale o critico. Quelli di sinistra, seppur negativi, appaiono relativamente più distesi o moderati. I subreddit neutrali hanno toni leggermente migliori rispetto ai subreddit di destra, ciò potrebbe indicare che siano luoghi di "scontro" tra utenti di ideologie diverse.

```
# Combino tutti i post in un solo dataframe
all_posts <- bind_rows(
  right_posts %>% mutate(group = "Destra"),
  left_posts %>% mutate(group = "Sinistra"),
  neutral_posts %>% mutate(group = "Neutro")
)

# Test di Kruskal-Wallis per confrontare il sentiment tra i gruppi
kruskal.test(sentiment ~ group, data = all_posts)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  sentiment by group
## Kruskal-Wallis chi-squared = 51.903, df = 2, p-value = 5.363e-12
```

```
# Confronti a coppie se il test è significativo
pairwise.wilcox.test(all_posts$sentiment, all_posts$group, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  all_posts$sentiment and all_posts$group
##
##          Destra  Neutro
## Neutro   1       -
## Sinistra 9.6e-10 1.3e-08
##
## P value adjustment method: bonferroni
```

I risultati del test di Wilcoxon corretti attraverso il metodo Bonferroni indicano come non vi siamo differenze significative tra il sentiment dei subreddit di destra e quelli neutrali, mentre la differenza tra subreddit di sinistra e gli altri due risulta essere significativa.

LDA

```
right_corpus <- Corpus(VectorSource(right_posts$clean_text))
left_corpus <- Corpus(VectorSource(left_posts$clean_text))
neutral_corpus <- Corpus(VectorSource(neutral_posts$clean_text))
dtm_right <- DocumentTermMatrix(right_corpus)
dtm_left <- DocumentTermMatrix(left_corpus)
dtm_neutral <- DocumentTermMatrix(neutral_corpus)

#tengo solo termini più frequenti
dtm_right <- removeSparseTerms(dtm_right, 0.99)
dtm_left <- removeSparseTerms(dtm_left, 0.99)
dtm_neutral <- removeSparseTerms(dtm_neutral, 0.99)

#tengo solo righe non vuote
dtm_right_matrix <- as.matrix(dtm_right)
dtm_left_matrix <- as.matrix(dtm_left)
dtm_neutral_matrix <- as.matrix(dtm_neutral)
row_sums_right <- rowSums(dtm_right_matrix)
row_sums_left <- rowSums(dtm_left_matrix)
row_sums_neutral <- rowSums(dtm_neutral_matrix)
dtm_right <- dtm_right[row_sums_right > 0, ]
dtm_left <- dtm_left[row_sums_left > 0, ]
dtm_neutral <- dtm_neutral[row_sums_neutral > 0, ]
```

Per ciascun gruppo di subreddit viene creato un corpus testuale distinto, dopodichè i vari testi sono stati trasformati in una DTM, in cui ogni riga rappresenta un post e ogni colonna un termine.

Per ridurre la dimensionalità e rimuovere i termini troppo rari si è applicata una soglia di sparsezza di 0,99.

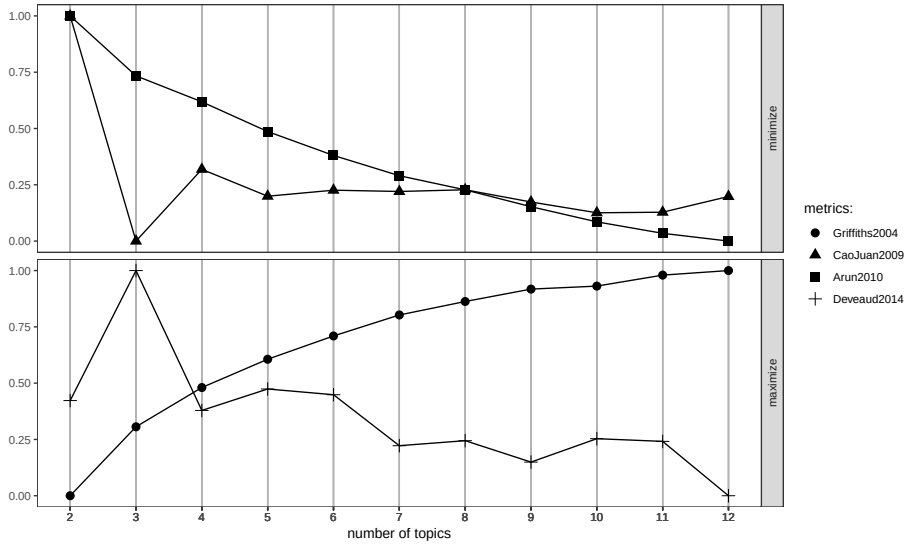
Successivamente sono stati esclusi i post che non contenevano alcun termine dopo la rimozione dei termini rari.

Per identificare il numero di topic ottimale vengono utilizzare quattro metriche:

- Griffiths2004 (scelta ottimale:massimo)
- CaoJuan2009(scelta ottimale:minimo)
- Arun2010(scelta ottimale:minimo)
- Deveaud2014(scelta ottimale:massimo)

```
optimal_right <- find_optimal_topics(dtm_right, k_min = 2, k_max = 12)
```

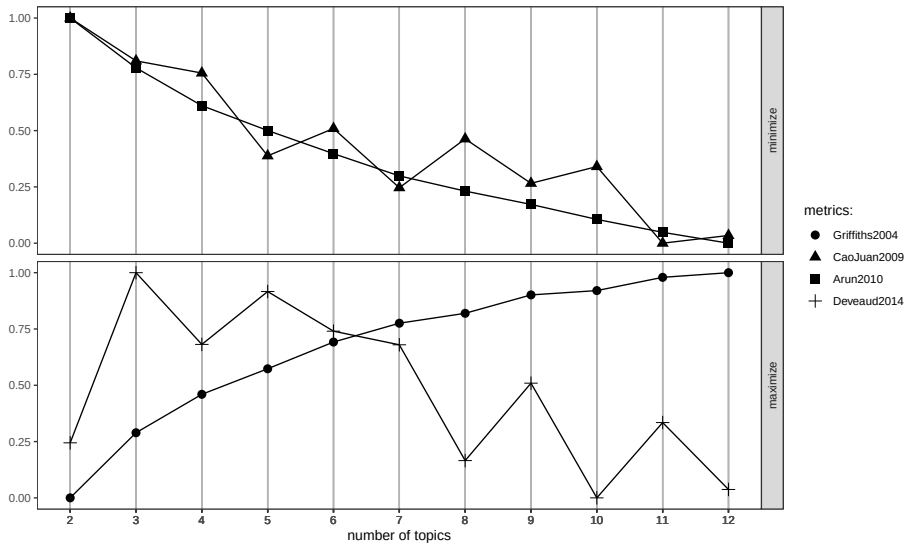
```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```



Gli indicatori mostrano un'andamento parzialmente discordante, tuttavia 3 topic sembra essere un compresso ottimale, essendo il minimo secondo CaoJuan2009 e il massimo secondo Deveaud2014.

```
optimal_left <- find_optimal_topics(dtm_left, k_min = 2, k_max = 12)
```

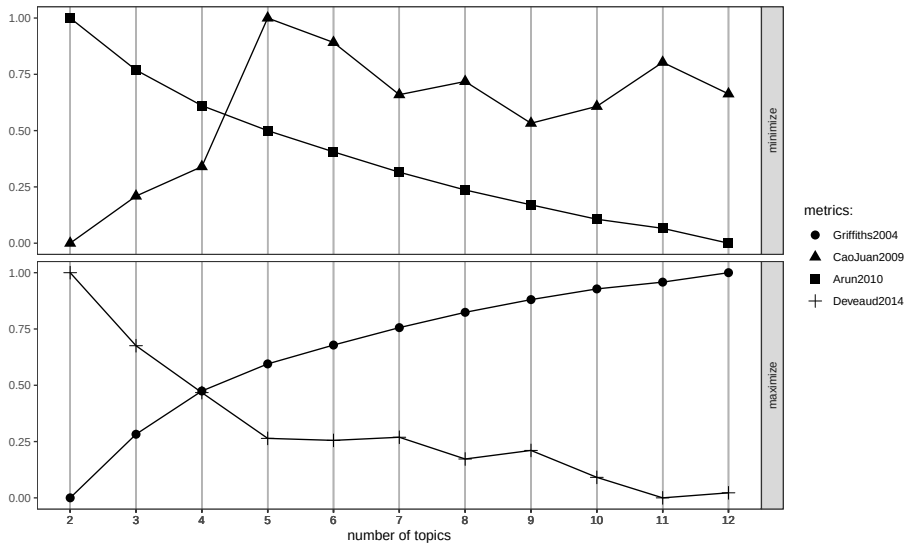
```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```



Anche in questo caso i metodi di selezione del numero ottimale di topic mostrano risultani divergenti, tuttavia il metodo Deveaud2014 mostra un picco elevato per 5 topic e i valori di CaoJuan2009 e Arun2010 sono relativamente bassi a 5 e 6 topic. Sulla base di questi elementi si è scelto di utilizzare 5 topic per i subreddit di sinistra.

```
optimal_neutral <- find_optimal_topics(dtm_neutral, k_min = 2, k_max = 12)
```

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```



Gli indicatori forniscono informazioni contrastanti, alcuni indicatori suggeriscono 2 topic, altri 10/12. Si è scelto di usarne 4 per avere un compromesso tra l'interpretabilità e distinzione semantica.

La differenze nel numero ottimale di topic forniscono delle informazioni riguardanti le caratteristiche dei vari topic, in particolare la presenza di più topic può indicare una maggiore varietà di argomenti o discussioni meno focalizzate. Di contro, la presenza di meno topic suggerisce una concentrazione tematica, con conversazioni dominate da pochi temi ricorrenti. I subreddit neutrali sembrano collocarsi in una posizione intermedia, suggerendo discussioni meno polarizzate ma non del tutto eterogenee.

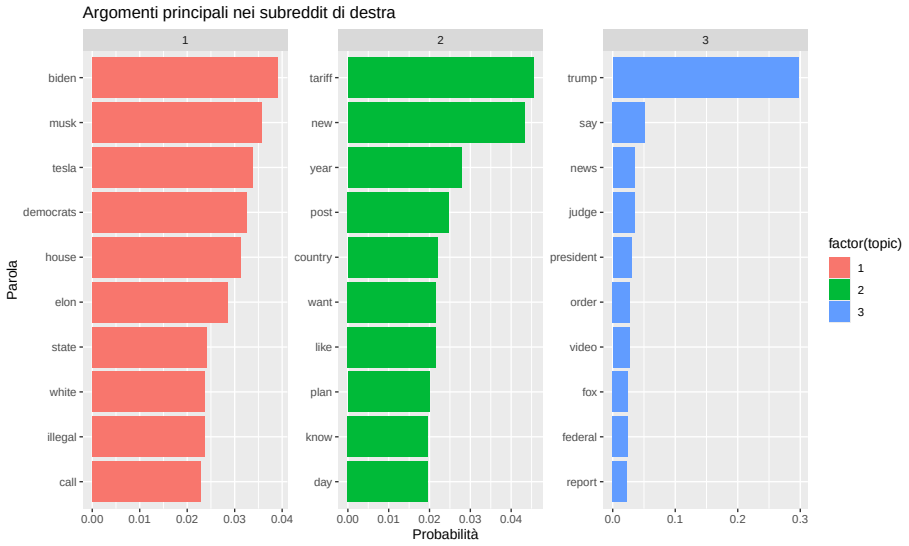
```
#LDA per le varie tipologie di post
lda_right <- LDA(dtm_right, k = 3, method="Gibbs",
                 control = list(seed = 1123,
                                burnin=1000, iter=1000))
lda_left <- LDA(dtm_left, k = 5, method="Gibbs",
                control = list(seed = 1123,
                               burnin=1000, iter=1000))
lda_neutral <- LDA(dtm_neutral, k = 4, method="Gibbs",
                   control = list(seed = 1123,
                                  burnin=1000, iter=1000))
```

E' stata applicata l'analisi dei topic tramite LDA per identificare i principali temi trattati nei post di ciascun gruppo di subreddit. In tutti i casi è stato utilizzato il metodo di campionamento "Gibbs sampling", con una fase di "burn-in" e un numero di interazioni sufficiente a garantire la convergenza del modello.

E' stato inoltre fissato un seed per garantire la riproducibilità del risultato.

Parole più usate subreddit di destra

```
topics_right <- tidy(lda_right, matrix = "beta")
top_terms_right <- topics_right %>%
  group_by(topic) %>%
  slice_max(beta, n = 10, with_ties = FALSE) %>%
  ungroup() %>%
  arrange(topic, -beta)
ggplot(top_terms_right, aes(x=reorder(term, beta), y=beta, fill=factor(topic))) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(title="Argomenti principali nei subreddit di destra",
       x="Parola", y="Probabilità") +
  facet_wrap(~topic, scales = "free")
```



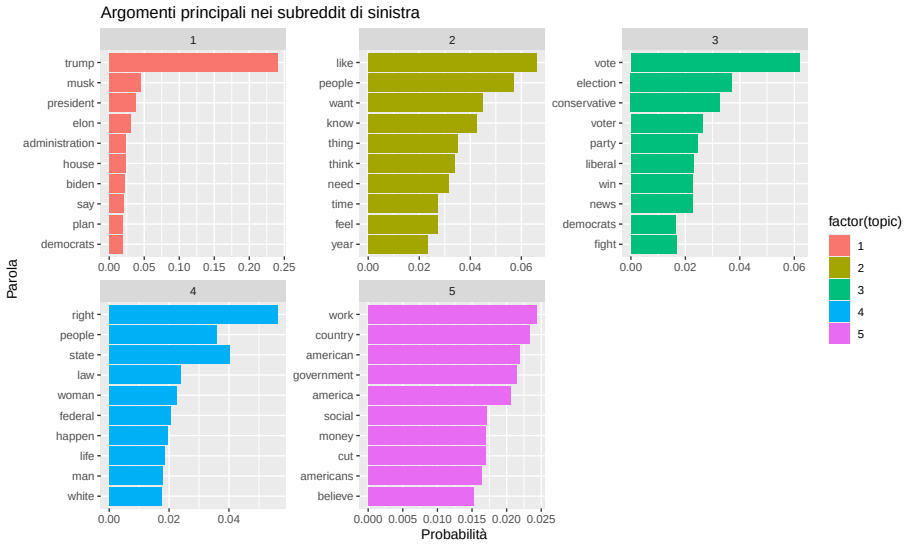
Il primo topic è caratterizzato da termini come “Biden”, “musk”, “tesla”, “democrats”, “elon”, “illegal”, “white” e “house”. Questo suggerisce un tema incentrato su figure poliche, istituzioni governative e personaggi o aziende legate alla tecnologia. Potrebbe rappresentare un dibattito critico sull’operato del governo Biden da parte di Musk, con accenni a temi identitari o controversi.

Il secondo topic contiene termini come “tariff”, “new”, “plan”, “year”, “country”, “want”, “day” e “know”. Possono indicare conversazioni su proposte politiche, sviluppo nazionale e aspettative verso il futuro economico del paese.

Il terzo topic, invece, è dominato da “Trump”, seguito da altri termini, questo topic è chiaramenti incentrato sulla figura di Trump, potrebbe rappresentare una narrativa difensiva o critica della copertura mediatica e giudiziaria su Trump.

Parole più usate subreddit di sinistra

```
topics_left <- tidy(lda_left, matrix = "beta")
top_terms_left <- topics_left %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
#grafico topic subreddit sinistra
ggplot(top_terms_left, aes(x=reorder(term, beta), y=beta, fill=factor(topic))) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(title="Argomenti principali nei subreddit di sinistra",
       x="Parola", y="Probabilità") +
  facet_wrap(~topic, scales = "free")
```



Il primo topic è dominato da “Trump”, insieme a “musk”, “elon”, “president”, “administration”, “biden”, “democrats”, questo topic sembra rappresentare riflessioni o critiche sull’amministrazione Trump e il suo impatto, il tono potrebbe essere critico verso l’attuale amministrazione o discutere la discontinuità rispetto a quella precedente.

I termini del secondo topic (“like”, “people”, “want”, “know”, “think”, “feel”, “need”, “time”) rappresentano delle discussioni più informali, con esperienze o percezioni del presente.

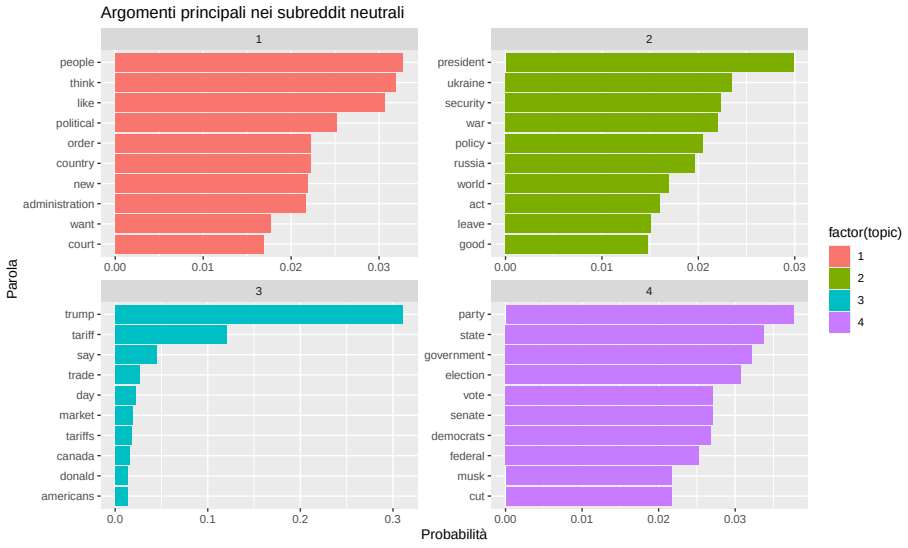
Il terzo topic contiene dei termini(“vote”, “election”, “conservative”, “voter”, “party”, “liberal”, “news”, “win”, “democrats”) riferiti alla competizione elettorale, probabilmente si tratta di reazioni ai risultati.

Nel quarto topic parole come “right”, “law”, “woman”, “federal”, “white”, “life”, “man”, “state” possono riguardare discussioni su diritti individuali e collettivi. Il dibattito può essere incentrato sul diritto all’aborto.

Nel quinto topic, le parole(“work”, “country”, “american”, “government”, “money”, “social”, “cut”) indicano discussioni su spesa pubblica e tagli, riflette probabilmente critiche sui tagli effettuati dall’amministrazione Trump.

Parole più usate subreddit di neutrali

```
topics_neutral <- tidy(lda_neutral, matrix = "beta")
top_terms_neutral <- topics_neutral %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
#grafico topic subreddit neutrali
ggplot(top_terms_neutral, aes(x=reorder(term, beta), y=beta, fill=factor(topic))) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(title="Argomenti principali nei subreddit neutrali",
       x="Parola", y="Probabilità") +
  facet_wrap(~topic, scales = "free")
```



Il primo topic con le parole “people”, “think”, “like”, “political”, “country”, “order”, “administration”, “court”, “want” rappresenta discussioni riguardanti l’operato dell’amministrazione corrente e su temi di governance, il tono sembra

essere riflessivo e valutativo, con riferimenti sia soggettivi (“think”,“like”) che istituzionali(“administration”, “court”).

Il secondo topic riguarda la politica estera e il conflitto Russia-Ucraina, lo si può comprendere guardando le parole “president”, “ukraine”, “russia”, “war”, “security”, “policy”, “world”, “act”.

Il terzo topic è dominato da “Trump” e “tariff” con “trade”, “market”, “americans”, “canada”, “day”, è chiaramente incentrato sulle nuove politiche economiche di Trump con un focus sui dazi e le loro conseguenze.

Il quarto topic invece, viste le parole che lo compongono(“party”, “state”, “government”, “election”, “vote”, senate, “democrats”, “federal”, “cut”) riguarda il sistema politico e le elezioni.

I subreddit di sinistra tendono ad essere più focalizzati su figure pubbliche, sull’economia e sui media. I topic risultano spesso concentrati su narrazioni mediatiche, temi politici e politiche economiche.

Anche nei subreddit di sinistra compaiono i personaggi pubblici, ma emerge una maggiore varietà nei temi, i topic incluno i diritti cilivi, esperienze personali, percezioni soggettive e riflessioni sullo Stato.

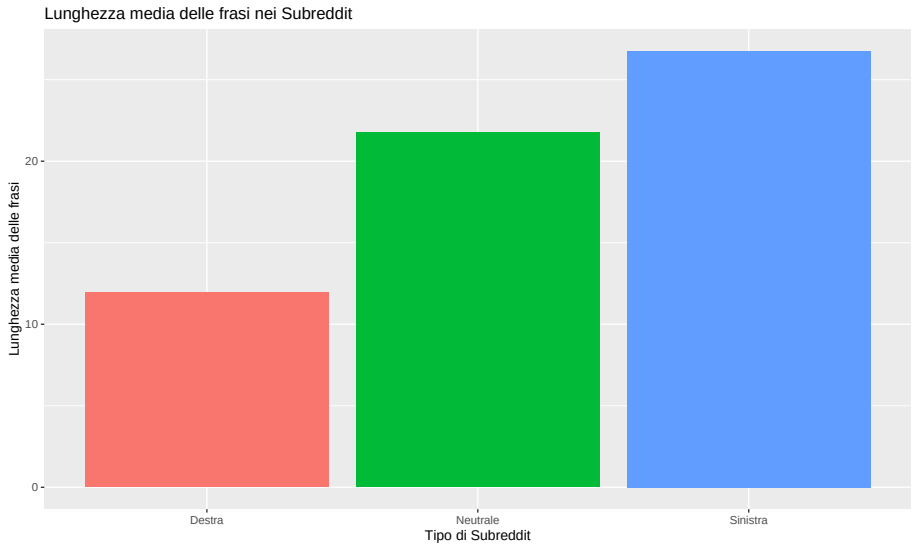
Nei subreddit neutrali è presente minore polarizzazione ideologica, i principali temi riguardano la politica estera, la struttura politica, percezioni generali sull’amministrazione e sul sistema giudiziario.

Complessità linguistica

Complessità linguistica post

Lunghezza media

```
right_posts$avg_sentence_length <- sapply(right_posts$clean_text, get_sentence_length)
left_posts$avg_sentence_length <- sapply(left_posts$clean_text, get_sentence_length)
neutral_posts$avg_sentence_length <- sapply(neutral_posts$clean_text, get_sentence_length)
avg_sentence_length <- data.frame(
  Subreddit = c("Destra","Sinistra", "Neutrale"),
  Avg_Sentence_Length = c(mean(right_posts$avg_sentence_length),
              mean(left_posts$avg_sentence_length),
              mean(neutral_posts$avg_sentence_length)))
ggplot(avg_sentence_length, aes(x=Subreddit, y=Avg_Sentence_Length, fill=Subreddit)) +
  geom_bar(stat="identity", show.legend = FALSE) +
  labs(title="Lunghezza media delle frasi nei Subreddit",
       y="Lunghezza media delle frasi", x="Tipo di Subreddit")
```

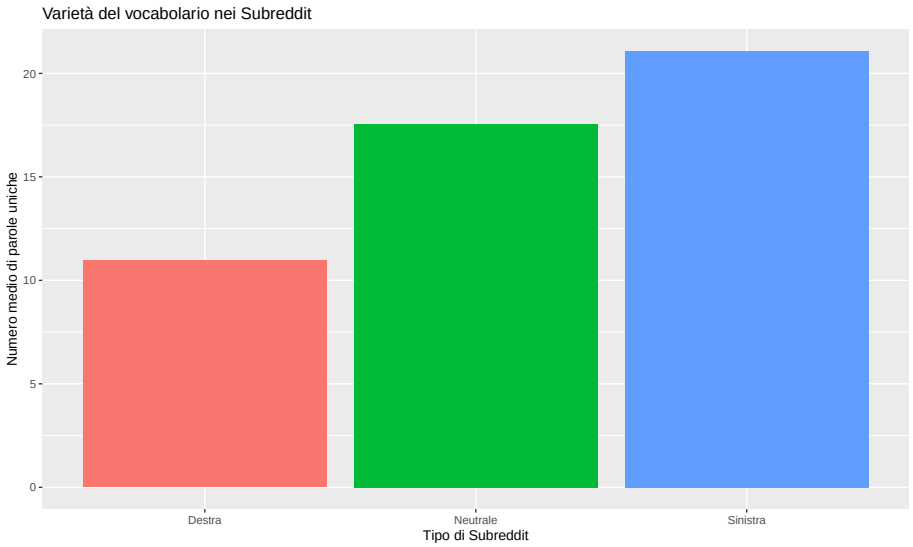


I post nei subreddit di sinistra mostrano una lunghezza media delle frasi nettamente maggiore, suggerendo un'articolazione maggiore. Al contrario i post nei subreddit di destra sono significativamente più concisi, con frasi più brevi e dirette.

I subreddit neutrali si collocano in una posizione intermedia, riflettendo un tono più moderato anche nella strutturra linguistica.

Parole uniche

```
right_posts$vocab_size <- sapply(right_posts$clean_text, get_vocab_size)
left_posts$vocab_size <- sapply(left_posts$clean_text, get_vocab_size)
neutral_posts$vocab_size <- sapply(neutral_posts$clean_text, get_vocab_size)
vocab_summary <- data.frame(
  Subreddit = c("Destra","Sinistra", "Neutrale"),
  Vocab_Size = c(mean(right_posts$vocab_size),
                 mean(left_posts$vocab_size),
                 mean(neutral_posts$vocab_size))
)
ggplot(vocab_summary, aes(x=Subreddit, y=Vocab_Size, fill=Subreddit)) +
  geom_bar(stat="identity", show.legend = FALSE) +
  labs(title="Varietà del vocabolario nei Subreddit",
       y="Numero medio di parole uniche", x="Tipo di Subreddit")
```



Anche in termini di ricchezza lessicale assoluta, i subreddit di sinistra mostrano una maggiore varietà linguistica, con un vocabolario quasi doppio rispetto a quello della destra.

I subreddit di destra risultano quindi più limitati nella diversità di termini impiegati, mentre quelli subreddit neutrali si collocano in una posizione intermedia, suggerendo una complessità espressiva più bilanciata.

TTR

```
combined_data <- merge(avg_sentence_length, vocab_summary, by = "Subreddit")
combined_data$TTR <- combined_data$Vocab_Size/ combined_data$Avg_Sentence_Length
combined_data <- combined_data[order(combined_data$TTR, decreasing = TRUE), ]
combined_data
```

```
##   Subreddit Avg_Sentence_Length Vocab_Size       TTR
## 1    Destra            11.95381   10.95811 0.9167041
## 2  Neutrale            21.76545   17.54530 0.8061078
## 3  Sinistra            26.75270   21.07866 0.7879079
```

Attraverso il TTR(Type-Token Ratio) possiamo osservare come il quadro si inverte: i subreddit di destra, pur utilizzando frasi più brevi e un vocabolario più limitato in termini assoluti, mostrano una maggiore varietà lessicale relativa, con un TTR più alto. Al contrario i subreddit di sinistra tengono a ripetere di più le parole, nonostante l'uso di un vocabolario più ampio e frasi più articolate. I subreddit neutrali invece si collocano in una posizione intermedia, sia per lunghezza che varietà.

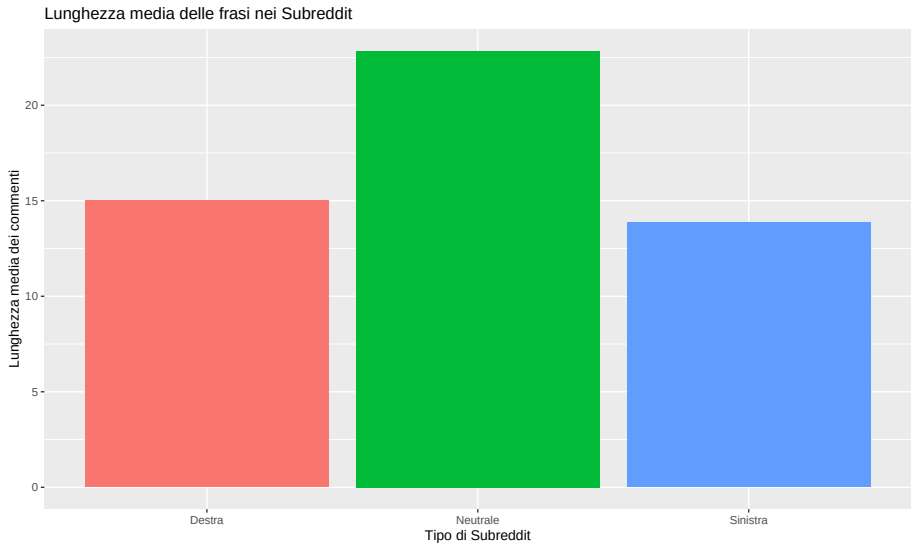
I tre gruppi, quindi, mostrano strategie linguistiche differenti nei post:

- Nei subreddit di sinistra viene usato un linguaggio più lungo e articolato, con molte parole uniche in valore assoluto, ma con maggiore ripetitività.
- I subreddit neutrali, si posizionano nel mezzo in ogni misura, coerente con l'idea di equilibrio
- I subreddit di destra hanno post brevi e concisi, con poche parole in assoluto, ma con grande varietà relativa. Presentano poca ridondanza, ciò indica uno stile più diretto ed efficace

Complessità lingustica commenti

Lunghezza media

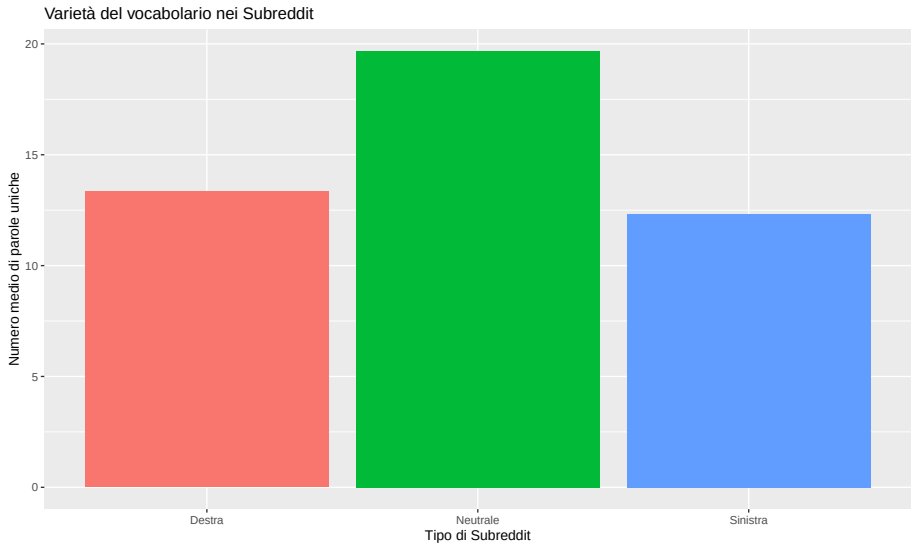
```
right_comments$avg_sentence_length <- sapply(right_comments$clean_body,
                                             get_sentence_length)
left_comments$avg_sentence_length <- sapply(left_comments$clean_body,
                                            get_sentence_length)
neutral_comments$avg_sentence_length <- sapply(neutral_comments$clean_body,
                                               get_sentence_length)
avg_sentence_length <- data.frame(
  Subreddit = c("Destra","Sinistra", "Neutrale"),
  Avg_Sentence_Length = c(mean(right_comments$avg_sentence_length),
                          mean(left_comments$avg_sentence_length),
                          mean(neutral_comments$avg_sentence_length))
)
ggplot(avg_sentence_length, aes(x=Subreddit, y=Avg_Sentence_Length, fill=Subreddit)) +
  geom_bar(stat="identity", show.legend = FALSE) +
  labs(title="Lunghezza media delle frasi nei Subreddit",
       y="Lunghezza media dei commenti", x="Tipo di Subreddit")
```



Nei commenti, i subreddit neutrali presentano frasi mediamente più lunghe rispetto a quelli di destra e sinistra. Questo potrebbe indicare una comunicazione più articolata e riflessiva, coerente con un contesto non polarizzato, dove si cerca di spiegare o approfondire. Al contrario, i subreddit di sinistra e destra mostrano frasi più brevi, suggerendo uno stile comunicativo più diretto, tipiche di interazioni a "botta e risposta", dove l'obiettivo è quello di replicare o affermare un punto.

Parole uniche

```
right_comments$vocab_size <- sapply(right_comments$clean_body, get_vocab_size)
left_comments$vocab_size <- sapply(left_comments$clean_body, get_vocab_size)
neutral_comments$vocab_size <- sapply(neutral_comments$clean_body, get_vocab_size)
vocab_summary <- data.frame(
  Subreddit = c("Destra","Sinistra", "Neutrale"),
  Vocab_Size = c(mean(right_comments$vocab_size),
                 mean(left_comments$vocab_size),
                 mean(neutral_comments$vocab_size))
)
ggplot(vocab_summary, aes(x=Subreddit, y=Vocab_Size, fill=Subreddit)) +
  geom_bar(stat="identity", show.legend = FALSE) +
  labs(title="Varietà del vocabolario nei Subreddit",
       y="Numero medio di parole uniche", x="Tipo di Subreddit")
```



I subreddit neutrali mostra una maggiore chiarezza lessicale rispetto ai subreddit di destra e di sinistra. Questo dato rafforza l'ipotesi che nei contesti meno polarizzati si tenda a impiegare un linguaggio più vario e articolato.

TTR

```
combined_data <- merge(avg_sentence_length, vocab_summary, by = "Subreddit")
combined_data$TTR <- combined_data$Vocab_Size/ combined_data$Avg_Sentence_Length
combined_data <- combined_data[order(combined_data$TTR, decreasing = TRUE), ]
combined_data
```

```
##   Subreddit Avg_Sentence_Length Vocab_Size       TTR
## 3  Sinistra            13.86077   12.33962 0.8902556
## 1    Destra            15.01438   13.35093 0.8892093
## 2  Neutrale            22.83513   19.67663 0.8616824
```

Osservando il TTR però il quadro si inverte leggeremente, i commenti nei subreddit di sinistra e di destra hanno una TTR più alta, quindi una maggiore varietà relativa del linguaggio, anche se usano meno parole in assoluto. I commenti nei subreddit neutrali, pur usando più parole diverse, tendono a ripetere di più.

Quindi, nei commenti il comportamento linguistico differisce dai post:

- Nei subreddit di sinistra sono presenti commenti brevi, ma con un linguaggio relativamente poco rindondante

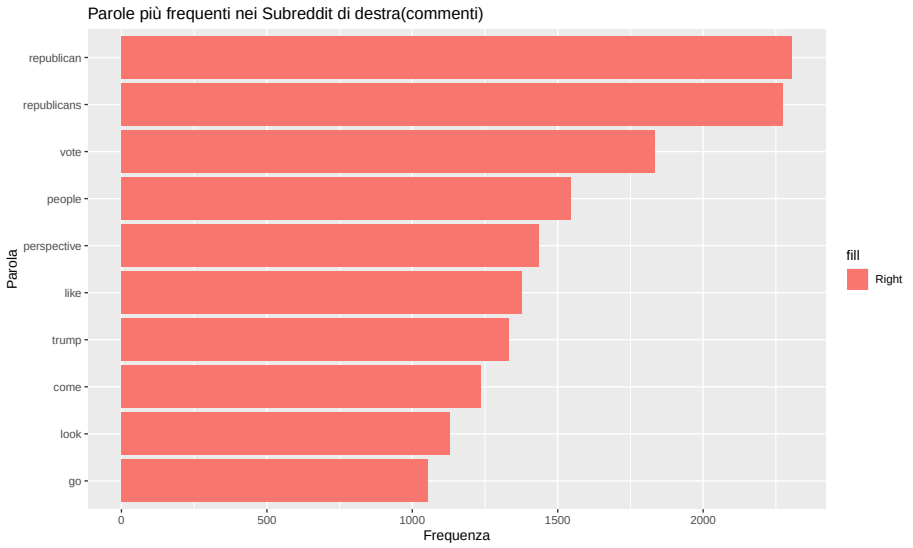
- I commenti nei subreddit di destra sono leggermente più lunghi rispetto a quelli nei subreddit di sinistra, ma sono anche leggermente più ripetitivi
- I commenti nei subreddit neutrali risultano essere più lunghi ma maggiormente ridondanti

Frequenza delle parole più comuni

```
right_words <- unlist(strsplit(right_comments$clean_body, " "))
left_words <- unlist(strsplit(left_comments$clean_body, " "))
neutral_words <- unlist(strsplit(neutral_comments$clean_body, " "))
right_word_freq <- as.data.frame(table(right_words))
left_word_freq <- as.data.frame(table(left_words))
neutral_word_freq <- as.data.frame(table(neutral_words))
right_word_freq <- right_word_freq %>% arrange(desc(Freq)) %>% head(10)
left_word_freq <- left_word_freq %>% arrange(desc(Freq)) %>% head(10)
neutral_word_freq <- neutral_word_freq %>% arrange(desc(Freq)) %>% head(10)
```

Si estraggono le parole da tutti i commenti dei tre gruppi, viene poi calcolata la frequenza assoluta delle parole e successivamente stilata una classifica, attraverso questo si identificano i termini maggiormente utilizzati nei commenti di ogni categoria.

```
ggplot(right_word_freq, aes(x=reorder(right_words, Freq), y=Freq, fill="Right")) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(title="Parole più frequenti nei Subreddit di destra(commenti)",
       x="Parola", y="Frequenza")
```

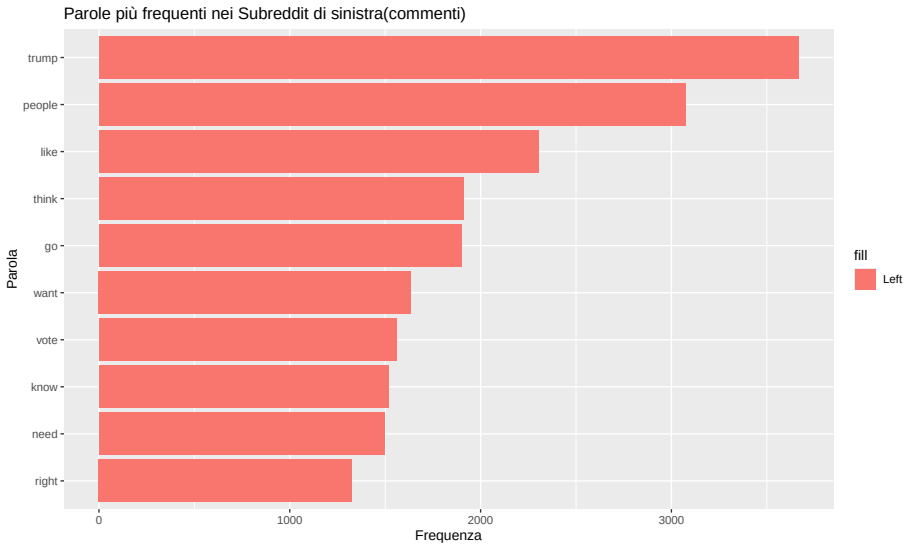


Il vocabolario dei commenti nei subreddit di destra è incentrato sull'identità e azione politica("republican"). "vote","Trump" suggerisce discussioni fortemente orientate all'ideologia e alle elezioni.

L'uso di parole come "perspective", "people", "like", "look" suggerisce uno stile che mescola la retorica emotiva con l'opinione personale.

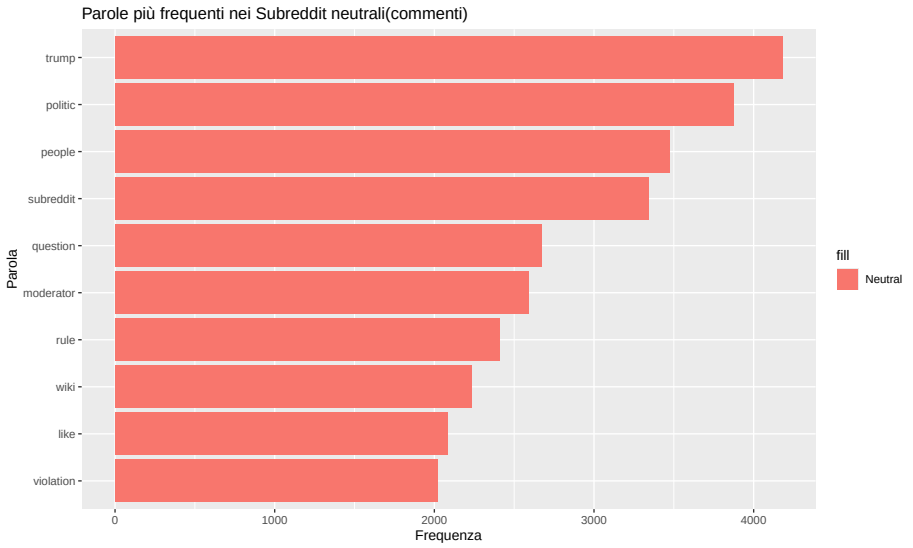
Il linguaggio sembra quindi piuttosto semplice e verbale, pochi termini astratti e molte parole comuni, questo supporta quanto detto in precedenza, frasi brevi ma con alta densità di significato.

```
ggplot(left_word_freq, aes(x=reorder(left_words, Freq), y=Freq, fill="Left")) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(title="Parole più frequenti nei Subreddit di sinistra(commenti)",
       x="Parola", y="Frequenza")
```



Il vocabolario dei commenti nei subreddit di sinistra è caratterizzato dalla centralità del "nemico" ideologico, probabilmente con toni critici e ossessivi. Parole come "want", "need", e "right" indicano un linguaggio normativo e valoriale, focalizzato su ciò che si desidera, si ritiene giusto o necessario. La presenza di verbi riflessivi("think","know") denotato un tono discorsivo, con spazio per opinioni personali.

```
ggplot(neutral_word_freq, aes(x=reorder(neutral_words, Freq), y=Freq, fill="Neutral")) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(title="Parole più frequenti nei Subreddit neutrali(commenti)",
       x="Parola", y="Frequenza")
```



Nel vocabolario dei commenti dei subreddit neutrali, troviamo sia termini politici rilevanti come "Trump" e "politic", ma anche molte parole che non riguardano direttamente contenuti politici.

Termini come "subreddit", "moderator", "rule", "wiki", e "violation" indicano un focus sulle regole della piattaforma e sulla gestione della comunità.

L'elevata frequenza di parole come "question" e "people" suggerisce un contesto in cui prevalgono domande e chiarimenti.

Analisi delle emozioni

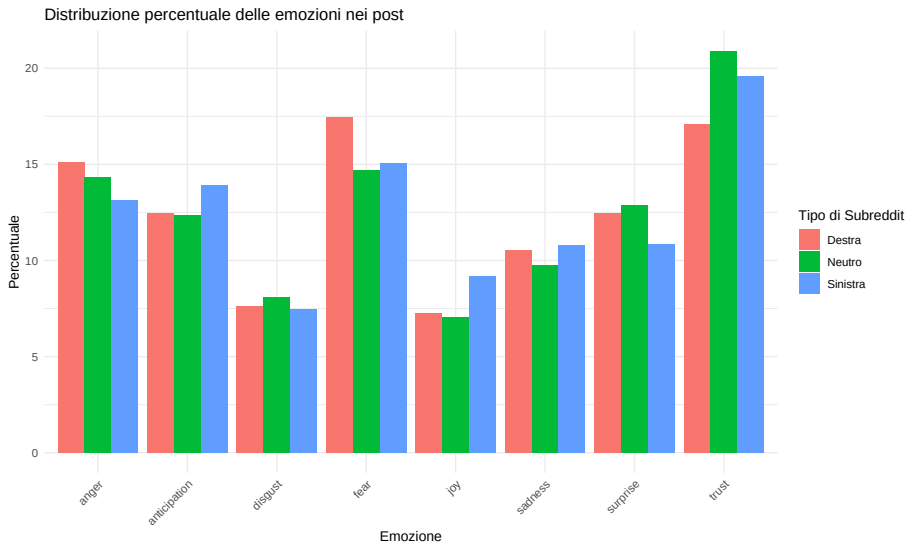
Analisi emozioni post

```
posts_tokens <- posts %>%
  select(clean_text, subreddit_type) %>%
  unnest_tokens(word, clean_text)
emotions_df <- get_nrc_sentiment(posts$clean_text)
posts_emotions <- cbind(posts, emotions_df)
avg_emotions <- posts_emotions %>%
  group_by(subreddit_type) %>%
  summarise(across(anger:trust, mean, na.rm = TRUE))
emotion_summary <- posts_emotions %>%
  group_by(subreddit_type) %>%
  summarise(across(anger:trust, sum, na.rm = TRUE))
emotion_percentages <- emotion_summary %>%
  mutate(total = rowSums(across(anger:trust))) %>%
```

```
  mutate(across(anger:trust, ~ .x / total * 100)) %>%
  select(-total)
emotion_long <- emotion_percentages %>%
  pivot_longer(cols = anger:trust, names_to = "emotion", values_to = "percent")
```

E' stato applicato il dizionario NRC ai post, associando a ciascun testo 10 emozioni. I valori sono stati poi normalizzati in percentuale sul totale di emozioni per ciascun tipo di subreddit.

```
ggplot(emotion_long, aes(x = emotion, y = percent, fill = subreddit_type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribuzione percentuale delle emozioni nei post",
       x = "Emozione", y = "Percentuale",
       fill = "Tipo di Subreddit") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



- Anger: più elevata nei subreddit di destra, coerente con un tono spesso polemico o indignato
- Anticipation: Maggiore nei subreddit di sinistra, forse legato a richieste di cambiamento o aspettative di progresso
- Disgut: Leggermente più alto nei subreddit neutrali, probabilmente per il frequente uso di parole normative o sanzionatorie
- Fear: Più alta nei subreddit di destra, segno di preoccupazione o percezione di minaccia(es. immigrazione)

- Joy: Chiaramente più alta nei subreddit di sinistra, riflettendo un tono più positivo o speranzoso
- Sadness: Più elevata nei subreddit di sinistra, forse legato a tematiche sociali, ingiustizia o empatia
- Surprise: Più alta nei subreddit neutrali, coerente con un tono più esplorativo o informativo
- Trust: massima nei subreddit neutrali, segno di un tono più pacato, regolato e meno confrittuale, forse perchè i subreddit neutrali hanno una moderazione più severa.

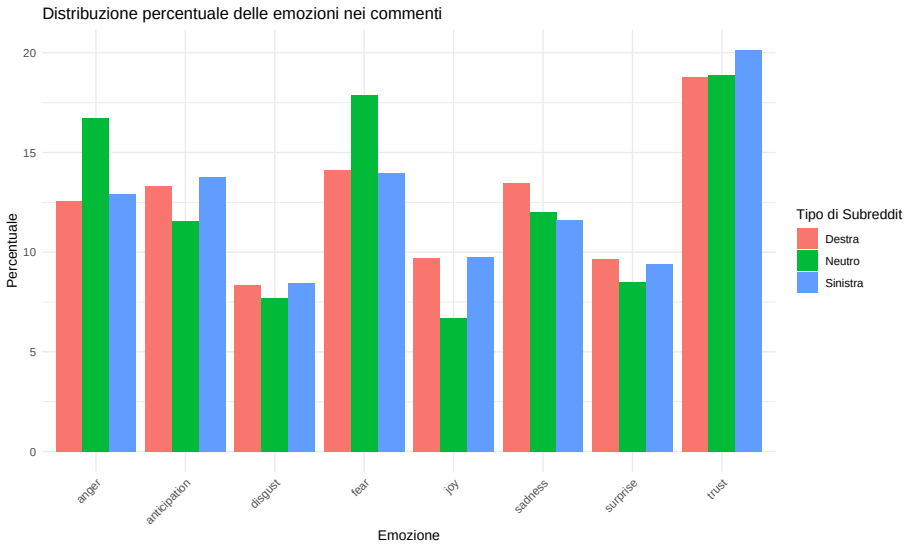
Calcolo emozioni commenti

```
comments_tokens <- comments %>%
  select(comment_id, clean_body, subreddit_type) %>%
  unnest_tokens(word, clean_body)
emotions_df <- get_nrc_sentiment(comments$body)
comments_emotions <- cbind(comments, emotions_df)

avg_emotions <- comments_emotions %>%
  group_by(subreddit_type) %>%
  summarise(across(anger:trust, ~ mean(.x, na.rm = TRUE)))

emotion_summary <- comments_emotions %>%
  group_by(subreddit_type) %>%
  summarise(across(anger:trust, sum)) %>%
  pivot_longer(cols = anger:trust, names_to = "emotion", values_to = "count")
emotion_percentages <- comments_emotions %>%
  group_by(subreddit_type) %>%
  summarise(across(anger:trust, sum, na.rm = TRUE)) %>%
  mutate(total = rowSums(across(anger:trust))) %>%
  mutate(across(anger:trust, ~ .x / total * 100)) %>%
  select(-total)

emotion_long <- emotion_percentages %>%
  pivot_longer(cols = anger:trust, names_to = "emotion", values_to = "percent")
```

Anche per analizzare il contenuto emotivo dei commenti è stato utilizzato il dizionario NRC.

```
ggplot(emotion_long, aes(x = emotion, y = percent, fill = subreddit_type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribuzione percentuale delle emozioni nei commenti",
```

```
      x = "Emozione", y = "Percentuale",
      fill = "Tipo di Subreddit") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



- Anger: più alta nei commenti dei subreddit neutrali, forse per lamentele, disaccordi o discussioni critiche su regole o contenuti.
- Anticipation: maggiore nei commenti dei subreddit di sinistra, suggerisce aspettative, proposte o progettualità; minima nei neutrali, coerente con uno stile più informativo o moderato.
- Disgust: le differenze sono modeste, ma i subreddit neutrali mostrano il valore più basso, coerente con un tono più regolato e impersonale.
- Fear: sorprendentemente, è più elevata nei commenti neutrali, il che potrebbe riflettere preoccupazioni generali o uno stile discorsivo più riflessivo.
- Joy: è massima nei commenti di sinistra, forse legata a toni più positivi o a discorsi su diritti, progresso o vittorie politiche.
- Sadness: i commenti dei subreddit di destra mostrano il livello più alto di tristezza, potenzialmente legato a sentimenti di perdita, ingiustizia o declino percepito.
- Surprise: le differenze sono contenute, ma i commenti nei subreddit neutrali appaiono leggermente meno sorpresi, coerenti con un linguaggio più analitico o normativo.

- Trust: i commenti nei subreddit di sinistra presentano il livello più alto di fiducia, suggerendo un discorso più coeso, orientato alla comunità o agli ideali.

Ponti narrativi

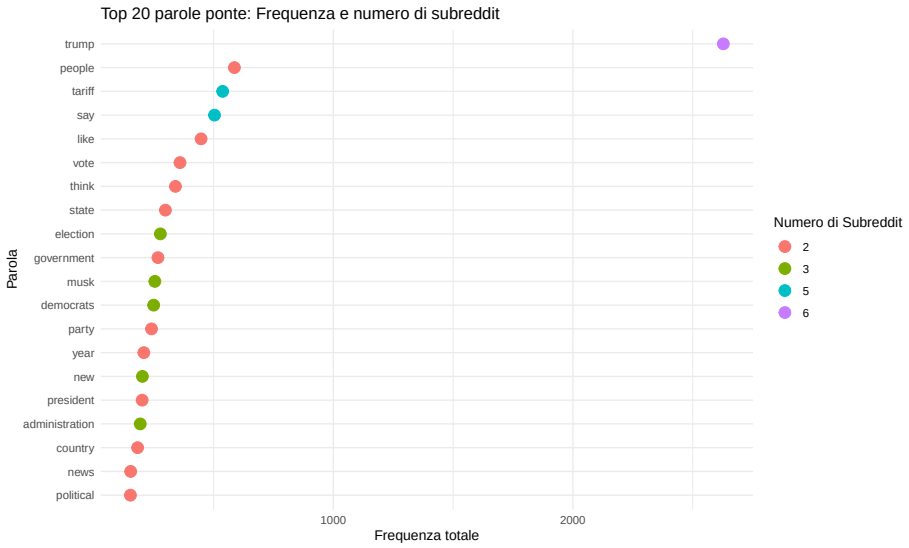
Ponti narrativi post

```
posts_tokens <- posts %>%
  select(post_id, subreddit, clean_text) %>%
  unnest_tokens(word, clean_text) %>%
  filter(nchar(word) > 2)
posts_tokens <- posts_tokens %>% filter(nchar(word) > 2)
word_counts <- posts_tokens %>%
  group_by(subreddit, word) %>%
  summarise(n = n(), .groups = "drop")
word_counts <- word_counts %>% filter(n >= 50)
word_groups <- word_counts %>%
  group_by(word) %>%
  summarise(
    subreddits = n_distinct(subreddit),
    total = sum(n),
    .groups = "drop"
  ) %>%
  arrange(desc(subreddits), desc(total))
bridge_words <- word_groups %>% filter(subreddits >= 2)
top_bridge <- bridge_words %>% top_n(20, total) %>% arrange(desc(total))
```

Quest'analisi individua le parole più trasversali tra i subreddit, ovvero i termini che ricorrono in più comunità politiche e che quindi possono essere considerate delle "parole-ponte". Il fatto che alcune parole compaiano in almeno due subreddit suggerisce che, nonostante le differenze ideologiche, esista un lessico tematico condiviso. Le parole-ponte con frequenza più alta non rappresentano solo un vocabolario comune, ma possono costituire dei nuclei semantici attorno a cui si articolano narrative contrapposte.

```
ggplot(top_bridge, aes(x = reorder(word, total),
                       y = total, color = as.factor(subreddits))) +
  geom_point(size = 4) +
  coord_flip() +
  labs(title = "Top 20 parole ponte: Frequenza e numero di subreddit",

       x = "Parola",
```

```
       y = "Frequenza totale",
       color = "Numero di Subreddit") +
  theme_minimal()
```



"Trump" è chiaramente la parola-ponte più trasversale. La sua presenza in tutti i subreddit sottolinea il ruolo centrale di Trump nel dibattito politico su reddit, sia come oggetto di sostegno che di critica. Funziona da nucleo narrativo condiviso, intorno al quale si articolano diverse prospettive.

Le parole presenti in più di tre subreddit ("tariff", "say" in cinque, "election", "musk", "democrats", "new", "administration" in tre) indicano temi ad ampio sprettro, spesso legati a eventi o figure di rilievo che coinvolgono diverse comunità politiche. Il fatto che siano presenti in almeno metà dei subreddit mostra che esiste un terreno comune su cui si innestano interpretazioni differenti, ma su temi condivisi.

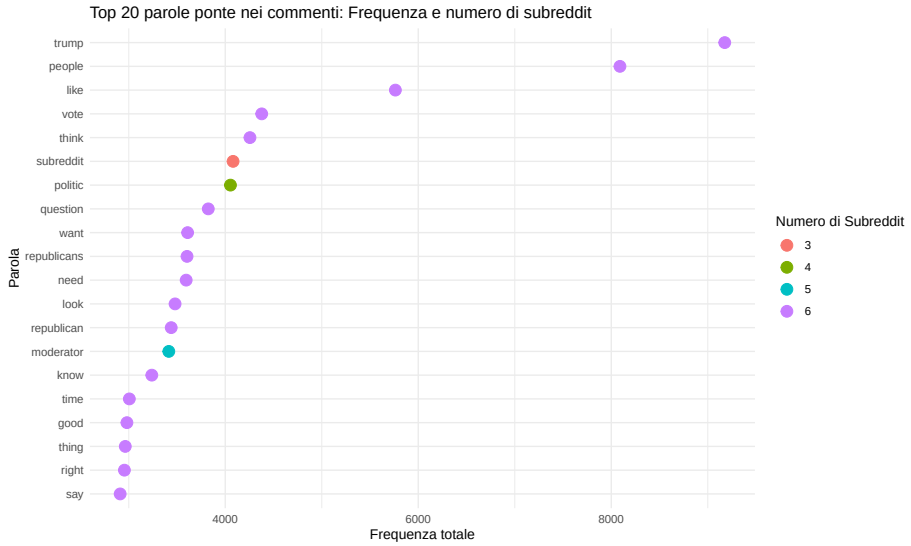
Le parole a bassa diffusione(presenti in 2 subreddit) sono parole molto generiche o settoriali. La loro presenza in soli due subreddit suggerisce fanno parte di narrazioni interne a determinati spazi ideologici.

Ponti narrativi commenti

```
comments_tokens <- comments %>%
  select(comment_id, subreddit, clean_body) %>%
  unnest_tokens(word, clean_body)
comments_tokens <- comments_tokens %>% filter(nchar(word) > 2)
word_counts <- comments_tokens %>%
```

```
  group_by(subreddit, word) %>%
  summarise(n = n(), .groups = "drop")
word_counts <- word_counts %>% filter(n >= 50)
word_groups <- word_counts %>%
  group_by(word) %>%
  summarise(
    subreddits = n_distinct(subreddit),
    total = sum(n),
    .groups = "drop"
  ) %>%
  arrange(desc(subreddits), desc(total))
bridge_words <- word_groups %>% filter(subreddits >= 2)
top_bridge <- bridge_words %>% top_n(20, total) %>% arrange(desc(total))
```

```
ggplot(top_bridge, aes(x = reorder(word, total),
                       y = total, color = as.factor(subreddits))) +
  geom_point(size = 4) +
  coord_flip() +
  labs(title = "Top 20 parole ponte nei commenti: Frequenza e numero di subreddit",

       x = "Parola",
       y = "Frequenza totale",
       color = "Numero di Subreddit") +
  theme_minimal()
```



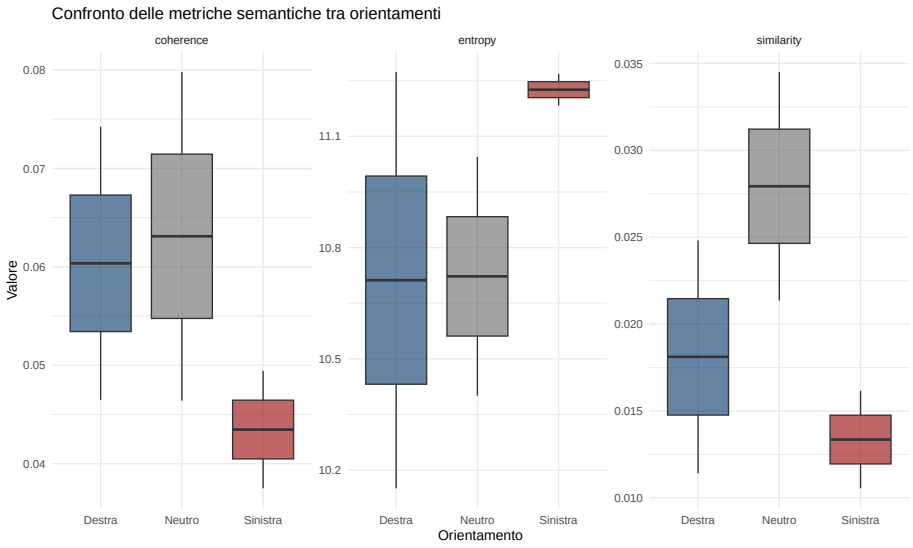
Nei commenti, a differenza dei post, un numero molto maggiore di parole appare in tutti e sei i subreddit. Questo suggerisce l'esistenza di un vocabolario

comune, fatto di termini generici, ricorrenti e funzionali al discorso, attorno a cui si costruiscono narrazioni politiche anche profondamente divergenti. Parole come “trump”, “people”, “vote”, “think”, “question”, “republicans”, “want” e “need” sono esempi di termini concettualmente aperti, che possono essere riempiti di significati diversi a seconda del contesto ideologico.

Analisi semantica post

```
sim_scores <- comments %>%
  group_by(subreddit) %>%
  summarise(similarity = get_cosine_similarity(clean_body))
thread_coherence <- comments %>%
  group_by(subreddit) %>%
  summarise(coherence = get_thread_coherence(cur_data()))
entropy_df <- comments %>%
  group_by(subreddit) %>%
  summarise(entropy = get_entropy(clean_body))
semantic_structure <- reduce(
  list(sim_scores, thread_coherence, entropy_df),
  left_join, by = "subreddit"
)
semantic_structure <- semantic_structure %>%
  mutate(tipo_subreddit = case_when(
    subreddit %in% c("Conservative", "Republican") ~ "Destra",
    subreddit %in% c("Liberal", "democrats") ~ "Sinistra",
    subreddit %in% c("PoliticalDiscussion", "politics") ~ "Neutro"
  ))
results_long <- semantic_structure %>%
  pivot_longer(cols = c(coherence, similarity, entropy),
               names_to = "metrica", values_to = "valore")
```

```
ggplot(results_long, aes(x = tipo_subreddit, y = valore, fill = tipo_subreddit)) +
  geom_boxplot(alpha = 0.6, outlier.shape = NA) +
  facet_wrap(~ metrica, scales = "free_y") +
  labs(title = "Confronto delle metriche semantiche tra orientamenti",
       x = "Orientamento", y = "Valore") +
  theme_minimal() +
  scale_fill_manual(values = c("Destra" = "#003366",
                               "Sinistra" = "#990000",
                               "Neutro" = "#666666")) +
  theme(legend.position = "none")
```



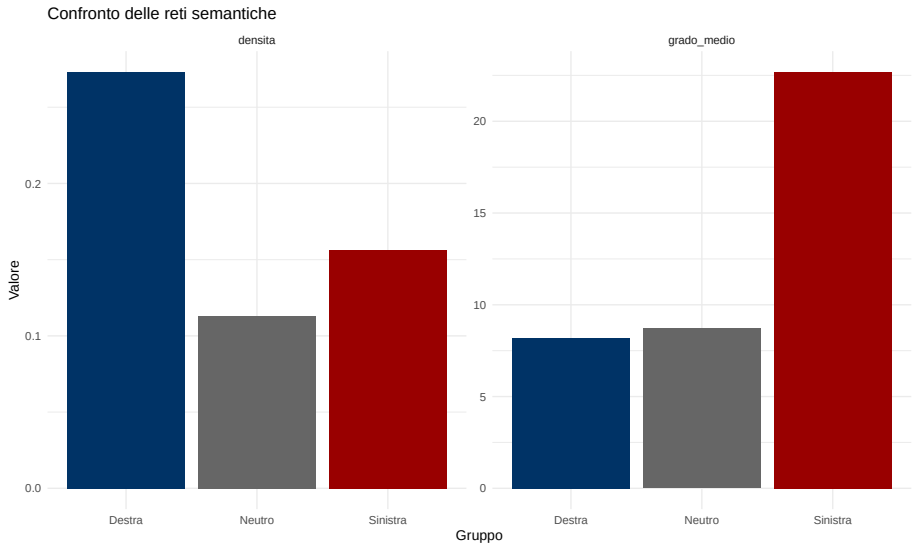
"Conservative" appare disgregato e diversificato, ha molte voci ma con poco allineamento argomentativo. L'entropia è alta, ciò indica la presenza di molte narrazioni frammentate. La comunità sempre polarizzata al suo interno. "Republican" al contrario, mostra una comunicazione coesa e orientata: alta coerenza e similarity, con bassa entropia. Probabile l'effetto di narrazioni dominanti, slogan ripetuti, o forte identificazione ideologia. Ciò può essere spiegato col fatto che "Republican" è il subreddit di un partito di destra, mentre "Conservative" è un subreddit che fa riferimento all'ideologia di destra.

Per quanto riguarda i subreddit di sinistra, entrambi mostrano una coerenza e similarity basse, indicando quindi la presenza di una comunicazione frammentata e spesso divergente. L'entropia è molto alta, e indicando quindi una forte varietà nei contenuti e nei punti di vista. La sinistra si distingue quindi per una comunicazione plurale, disomogenea, poco coordinata. Questo può riflettere una maggiore apertura al dissenso, ma anche una minore capacità di costruzione di messaggi coesi e mobilitanti.

Passando ai subreddit neutrali, "PoliticalDiscussion" è il subreddit più coeso e simile tra tutti, sembra essere un subreddit con regole discorsive forti e una moderazione particolarmente attiva. "Politics", invece, è meno strutturato ma comunque più ordinato rispetto ai subreddit politici più polarizzati. E' evidente che i subreddit neutrali agiscono come subreddit di discussione, l'alto livello di coerenza indica un certo grado di confronto, probabilmente favorito da linee guida più restittive o da una composizione più mista dell'utenza.

Analisi delle reti semantiche

```
tokens <- posts %>%
  select(post_id, subreddit_type, clean_text) %>%
  unnest_tokens(word, clean_text) %>%
  filter(str_length(word) > 2, !word %in% stop_words$word)
cooc_pairs <- tokens %>%
  group_by(subreddit_type) %>%
  pairwise_count(word, post_id, sort = TRUE, upper = FALSE)
metrics_by_group <- cooc_pairs %>%
  group_by(subreddit_type) %>%
  group_map(~ {
    metrics <- get_network_metrics(.x)
    metrics$subreddit_type <- unique(.x$subreddit_type)
    metrics
  }, .keep = TRUE) %>%
  bind_rows() %>%
  ungroup()
metrics_by_group %>%
  pivot_longer(cols = c("grado_medio", "densita"),
               names_to = "metrica", values_to = "valore") %>%
  ggplot(aes(x = subreddit_type, y = valore, fill = subreddit_type)) +
  geom_col(position = "dodge") +
  facet_wrap(~ metrica, scales = "free_y") +
  labs(title = "Confronto delle reti semantiche",
       x = "Gruppo", y = "Valore") +
  scale_fill_manual(values = c("Destra" = "#003366",
                               "Sinistra" = "#990000",
                               "Neutro" = "#666666")) +
  theme_minimal() +
  theme(legend.position = "none")
```



I subreddit di sinistra presentano il grado medio più alto, questo indica una rete molto più ricca e articolata sul piano semantico, dove molte parole ricorrono assieme in più post. Tuttavia, la densità non è la più alta, ciò indica che nonostante vi siano molti collegamenti, non tutte le parole sono connesse tra loro.

I subreddit di destra ha il grado medio più basso, ma la densità più alta, ciò suggerische che, pur avendo meno collegamenti per nodo, le parole tendono ad essere collegate in modo più equo, creando una rete semanticamente compatta ma meno centrale. Potrebbe indicare che i post sono più omogenei e i concetti chiavi si combinano tra loro più frequentemente.

Nei subreddit neutri invece si ha valori bassi sia di grado medio che di densità, le parole risultano essere meno collegate(rete più sparsa). Questo è coerente con una varietà tematica ampia e una minore polarizzazione.

Conclusioni

L'analisi evidenzia differenze sostanziali tra i gruppi di destra, sinistra e neutrali. I subreddit di destra presentano narrazioni ideologiche più focalizzate e coese, specie in comunità come "Republican", con linguaggi compatti e tematizzazioni ripetitive. Al contrario, i subreddit di sinistra mostrano una comunicazione più frammentata e plurale, caratterizzata da un'ampia varietà di argomenti e da una rete semantica molto articolata, che riflette una maggiore apertura a punti di vista diversi ma una minore coesione discorsiva complessiva. I subreddit neutrali, invece, emergono come spazi più strutturati e moderati, con topic equilibrati, reti semantiche meno dense, ma coese, e un linguaggio condiviso più razionale e dialogico.

Dal punto di vista emotivo, i subreddit di destra mostrano livelli elevati di tristezza e paura, ma anche una forte presenza di fiducia, segnalando una combinazione di preoccupazione e forte identificazione ideologica. I subreddit di sinistra esprimono maggiore gioia e fiducia, insieme a un'anticipazione proiettata verso il futuro, mentre quelli neutrali evidenziano maggiori livelli di rabbia e paura, probabilmente legati a discussioni critiche o riflessive su eventi e regole. L'analisi emotiva conferma così le dinamiche di coinvolgimento e il clima emotivo diverso che contraddistingue ciascun gruppo, contribuendo a definire la natura del dibattito politico su Reddit.

Infine, l'esame dei commenti rispetto ai post rivela come, sebbene i post tendano a costruire narrazioni fortemente identitarie e polarizzate, i commenti abbiano un vocabolario più condiviso e un tono più dialogico, indicando che lo scambio interattivo tra utenti si sviluppa su un terreno linguistico più comune e meno frammentato. Nel complesso, questo studio mostra come le dinamiche comunicative e emotive nei diversi subreddit riflettano non solo le divisioni politiche, ma anche differenti modalità di costruzione del discorso e partecipazione sociale.

Limiti e sviluppi futuri

Limiti

Il progetto presenta diversi limiti, a partire dalla rappresentatività dei dati, in quanto l'analisi è basata su sei subreddit specifici scelti per la loro identificabilità ideologica. Questi subreddit, però, non rappresentano l'intero dibattito politico su Reddit. Inoltre i dati raccolti includono solo gli ultimi 1000 post per ciascun subreddit, il che limita la copertura temporale e non consente di analizzare l'impatto di specifici eventi politici.

Inoltre l'analisi testuale si basa su metodi bag-of-words e dizionari statici, che non riesco a catturare l'ironia e il sarcasmo. Infine, la suddivisione in ideologie politiche è stata effettuata a priori e applicata uniformemente a tutti i post e commenti, tuttavia, all'interno di ogni comunità possono coesistere voci eterogenee e sottogruppi ideologici diversi.

Sviluppi futuri

Si potrebbe effettuare la raccolta dei dati in vicinanza di eventi politici e confrontare come cambi il discorso in base all'evento, ciò potrebbe offrire un quadro più ampio del discorso politico online. L'adozione di modelli di NLP più sofisticati(come Bert o GPT) permetterebbe di migliorare la comprensione semantica, consentendo di rilevare l'ironia o sentiment ambigui.

Sarebbe anche possibile unire l'analisi testuale a quella delle interazione tra utenti(upvote, menzioni, risposte) per poter studiare dinamiche di influenza, diffusione del contenuto e polarizzazione strutturale. Infine si potrebbe estendere il confronto ad altre piattaforme, esplorando quindi come la forma del medium influenzi il tipo di discorso politico.