CAPITOLO 5: Analisi commenti su X

Carichiamo il primo dataset: APIFY_1. È stato ottenuto tramite un software di scraping che ha permesso l'estrazione di gratuitamente di 841 tweet, trovati utilizzando come key-words le parole "AI", "artificial intellingence" e l'hastag "#AI debate".

5.1 Polarità

La polarità di un commento su Twitter (o di qualsiasi testo) si riferisce alla sua inclinazione emotiva, ovvero se il contenuto è positivo, negativo o neutro. È una metrica fondamentale per l'analisi del sentiment, utile per comprendere le opinioni degli utenti su un argomento o prodotto.

```
library(quanteda)
```

```
## Package version: 4.1.0
## Unicode version: 14.0
## ICU version: 71.1
```

```
## Parallel computing: disabled
```

```
## See https://quanteda.io for tutorials and examples.
```

```
library(qdap)
```

```
## Loading required package: qdapDictionaries
```

```
## Loading required package: qdapRegex
```

```
## Loading required package: qdapTools
```

```
## Loading required package: RColorBrewer
```

```
##
## Attaching package: 'qdap'
```

```
## The following objects are masked from 'package:base':
##
##     Filter, proportions
```

```
library(readxl)
tweet_df <- read_excel("DATASET_PULITO.xlsx")
```

```
## New names:
## • `` -> `...2`
## • `` -> `...14`
## • `` -> `...16`
## • `` -> `...18`
## • `` -> `...19`
## • `` -> `...20`
```

```r
polarity(tweet_df$text)
```

```
## Warning in polarity(tweet_df$text):
##   Some rows contain double punctuation.  Suggested use of `sentSplit` function.

##   all total.sentences total.words ave.polarity sd.polarity stan.mean.polarity
## 1 all             841       27118        0.163       0.223              0.732
```

Un valore di polarità di 0.163 indica una leggera tendenza positiva nei commenti analizzati su Twitter riguardo all'intelligenza artificiale. Questo suggerisce che, sebbene ci siano sia opinioni entusiastiche che timorose, il sentiment complessivo tende verso un atteggiamento più favorevole che negativo.

L'entusiasmo potrebbe derivare da commenti che elogiano i progressi dell'IA, la sua capacità di automatizzare compiti complessi e le opportunità che offre in diversi settori, come sanità, finanza e creatività. D'altro canto, il timore è spesso legato alle questioni di etica, privacy, perdita di posti di lavoro e possibili usi impropri della tecnologia.

5.2 Sentiment Analysis

L'analisi del sentiment è il processo automatizzato che esamina un testo per determinare l'opinione o l'emozione espressa dall'autore. Questo processo può essere visto come una classificazione binaria (positivo o negativo) o in classi multiple (es.felicità, tristezza,rabbia).

La libreria VADER (Valence Aware Dictionary and Sentiment Reasoner) è un algoritmo di analisi del sentiment progettato specificamente per testi brevi, come i tweet. Basato su un dizionario lessicale, VADER assegna punteggi di intensità alle parole per determinare il tono complessivo di un testo.

```r
library(vader)
sentiment_results <- vader_df(tweet_df$text)
head(sentiment_results)
```

```
##
text
## 1                                                                        89% of
employees believe that AI could improve at least half of their workload. Use these
8 AI productivity tools to give you super-human powers in 2024!
https://t.co/QMLIwfHc8t #ai #productivity #artificialintelligence #aitools
## 2                                              In the grand theatre of life, one
must always be ready for the unexpected entrance of beauty without rehearsal.
#aigirl #aimodels #virtualinfluencers #digitalfashion #artificialintelligence
#virtualbeauty #metaverse #futurevisions https://t.co/RKNPDzGbeb
## 3
Is Ethics Needed for Artificial Intelligence?  https://t.co/KyhvW0vg40
#ArtificialIntelligence
## 4
1863 letter warned against AI takeover!  A chillingly prescient call to halt
technological progress to prevent machine dominance resurfaces.  #AI
#ArtificialIntelligence #FutureofTech https://t.co/cBONiCA89Z
## 5
```

```
Online comments of tourist attractions combining artificial intelligence text
mining model and attention mechanism. https://t.co/BP8gum6YSO
#ArtificialIntelligence
## 6 Trend Alert! AI in digital marketing is shaping the future. Here's why your
business should embrace it.\r\n\r\n#digitalmarketing #aiinmarketing
#marketingtrends #aiforbusiness #futureofmarketing #digitaltransformation
#artificialintelligence #marketingstrategy #aiinnovation https://t.co/Ly0OAFhrfM
##
word_scores
## 1 {0, 0, 0, 0, 0, 0, 0, 1.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0}
## 2           {0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 1.5, 0, 0, 0, 0, 0, 2.8, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0}
## 3                                                                          {0,
0, 0, 0, 0, 2.1, 0, 0}
## 4                      {0, 0, -1.1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.8, 0, 0.1,
0, 0.8, 0, 0, 0, 0, 0}
## 5                                                 {0, 0, 0, 0, 1.8, 0, 0, 2.1, 0,
0, 0, 0, 0, 0, 0, 0}
## 6              {0, 1.2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.3, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0}
##   compound   pos   neu   neg but_count
## 1    0.493 0.093 0.907 0.000         0
## 2    0.852 0.271 0.729 0.000         0
## 3    0.477 0.307 0.693 0.000         0
## 4    0.439 0.221 0.701 0.078         0
## 5    0.710 0.296 0.704 0.000         0
## 6    0.585 0.161 0.839 0.000         0
```

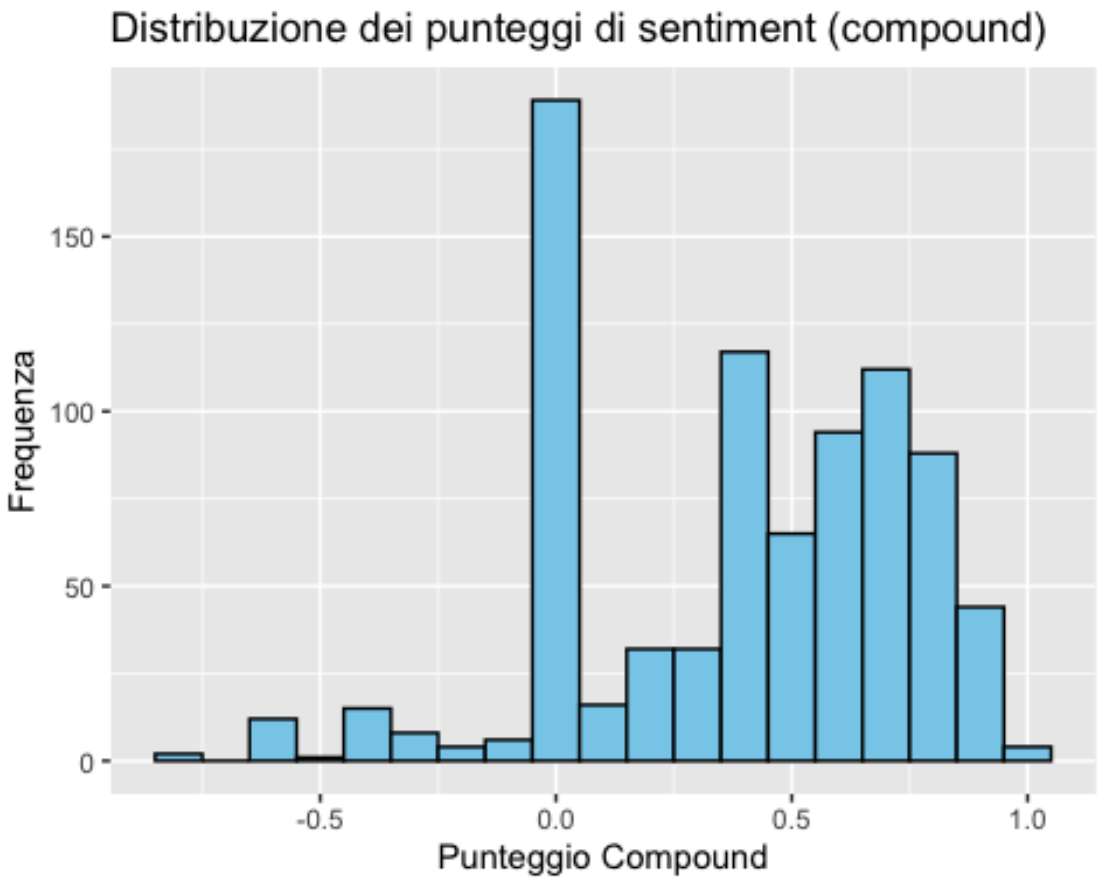
La funzione restituirà, per ogni tweet, diversi punteggi compound:

- un punteggio complessivo compreso tra -1 (molto negativo) e 1 (molto positivo)
- pos: La percentuale di sentiment positivo
- neu: La percentuale di sentiment neutro
- neg: La percentuale di sentiment negativo

Grafichiamo i risultati ottenuti

```
library(ggplot2)

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:qdapRegex':
##
##     %+%

ggplot(sentiment_results, aes(x = compound)) +
  geom_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +
  labs(title = "Distribuzione dei punteggi di sentiment (compound)",
```

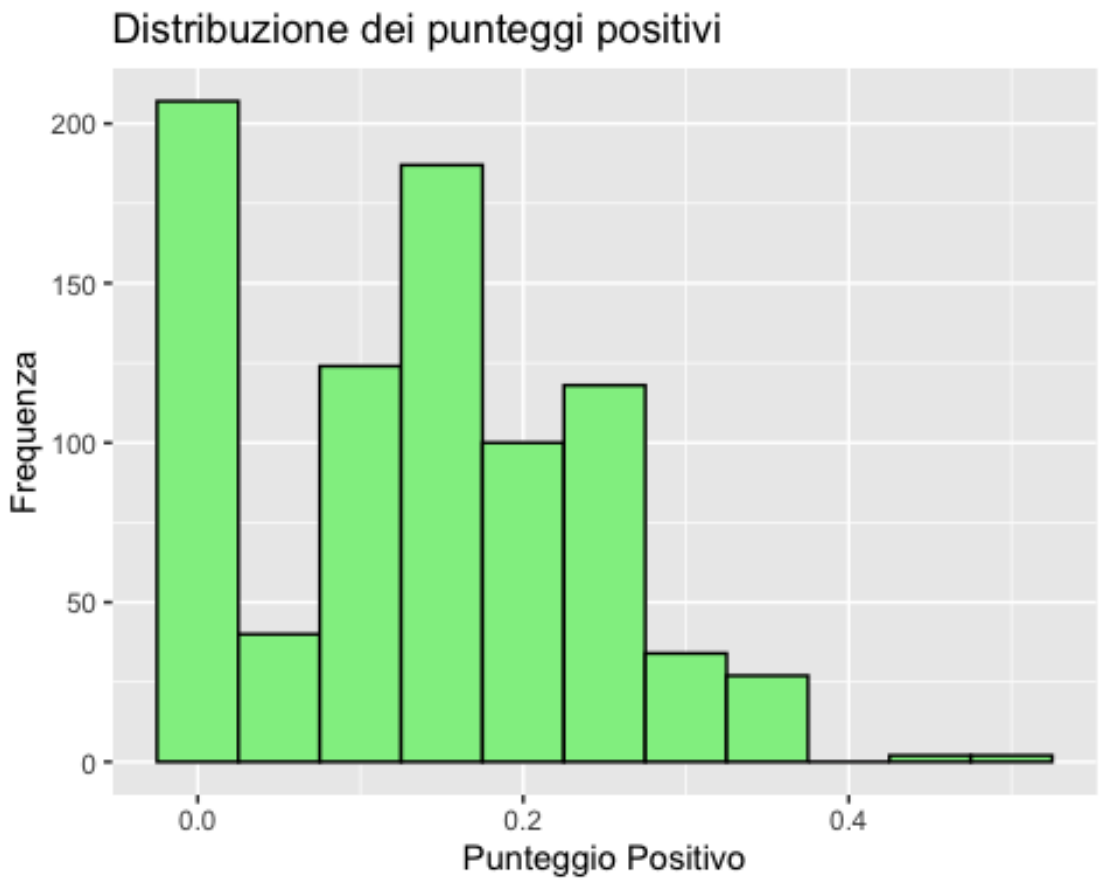
```
       x = "Punteggio Compound",
       y = "Frequenza")
```



L'istogramma rappresenta la distribuzione dei punteggi compound ottenuti con la libreria VADER per l'analisi del sentiment. La maggior parte dei valori si concentra intorno a punteggi neutri, con una leggera prevalenza di sentimenti positivi, evidenziata dalla presenza di punteggi compound intorno a 0.7-0.8. Questo suggerisce che il dataset analizzato contiene principalmente testi con un tono equilibrato o leggermente positivo, mentre le valutazioni estremamente negative o fortemente positive sono meno frequenti.

Grafichiamo ora la frequenza dei commenti positivi

```
ggplot(sentiment_results, aes(x = pos)) +
  geom_histogram(binwidth = 0.05, fill = "lightgreen", color = "black") +
  labs(title = "Distribuzione dei punteggi positivi",
       x = "Punteggio Positivo",
       y = "Frequenza")
```




```
library(quanteda)
mycorp <- corpus(tweet_df$text)
summary(mycorp)
```

```
## Corpus consisting of 841 documents, showing 100 documents:
##
##     Text Types Tokens Sentences
##    text1    33     35         3
##    text2    28     30         1
##    text3     9      9         2
##    text4    24     25         3
##    text5    17     17         1
##    text6    29     30         3
##    text7    36     38         2
##    text8    36     38         2
##    text9    36     38         2
##   text10    36     38         2
##   text11    36     38         2
##   text12    36     38         2
##   text13    36     38         2
##   text14    36     38         2
##   text15    36     38         2
##   text16    36     38         2
```

```
##   text17    18     22         1
##   text18    20     20         2
##   text19    17     17         1
##   text20    18     18         1
##   text21    14     14         1
##   text22    29     30         1
##   text23    20     20         1
##   text24    21     21         1
##   text25    33     35         3
##   text26    28     30         1
##   text27     9      9         2
##   text28    24     25         3
##   text29    17     17         1
##   text30    29     30         3
##   text31    36     38         2
##   text32    36     38         2
##   text33    36     38         2
##   text34    36     38         2
##   text35    36     38         2
##   text36    36     38         2
##   text37    36     38         2
##   text38    36     38         2
##   text39    36     38         2
##   text40    36     38         2
##   text41    18     22         1
##   text42    20     20         2
##   text43    17     17         1
##   text44    18     18         1
##   text45    34     39         2
##   text46    13     13         2
##   text47    22     22         1
##   text48    31     34         3
##   text49    15     16         1
##   text50    13     14         1
##   text51    20     20         1
##   text52    31     34         3
##   text53    31     34         3
##   text54    40     44         4
##   text55    40     44         4
##   text56    11     11         1
##   text57    31     34         3
##   text58    13     13         1
##   text59    31     34         3
##   text60    11     11         1
##   text61    17     18         1
##   text62    16     16         1
##   text63    15     15         1
##   text64    22     22         1
##   text65    26     26         1
##   text66    12     12         1
##   text67    22     22         1
```

```
##   text68    29     29         1
##   text69    24     24         1
##   text70    31     34         3
##   text71    36     44         4
##   text72    37     37         1
##   text73    21     21         1
##   text74    31     34         3
##   text75    22     23         1
##   text76    31     34         3
##   text77    23     26         3
##   text78    24     24         2
##   text79    25     28         3
##   text80    31     34         4
##   text81    36     39         3
##   text82    34     38         3
##   text83    28     28         2
##   text84    21     27         1
##   text85    34     34         2
##   text86    32     33         1
##   text87    29     33         2
##   text88    33     34         1
##   text89    28     29         2
##   text90    18     18         1
##   text91    29     33         2
##   text92    23     24         4
##   text93    22     22         1
##   text94    29     33         2
##   text95    29     33         2
##   text96    23     23         2
##   text97    19     19         1
##   text98    29     33         2
##   text99    29     33         2
##  text100    38     42         4
```

```
toks <- tokens(mycorp)
mycorp1<-dfm(toks, tolower=T)
tag_dfm <- dfm_select(mycorp1, pattern = ("#*"))
```

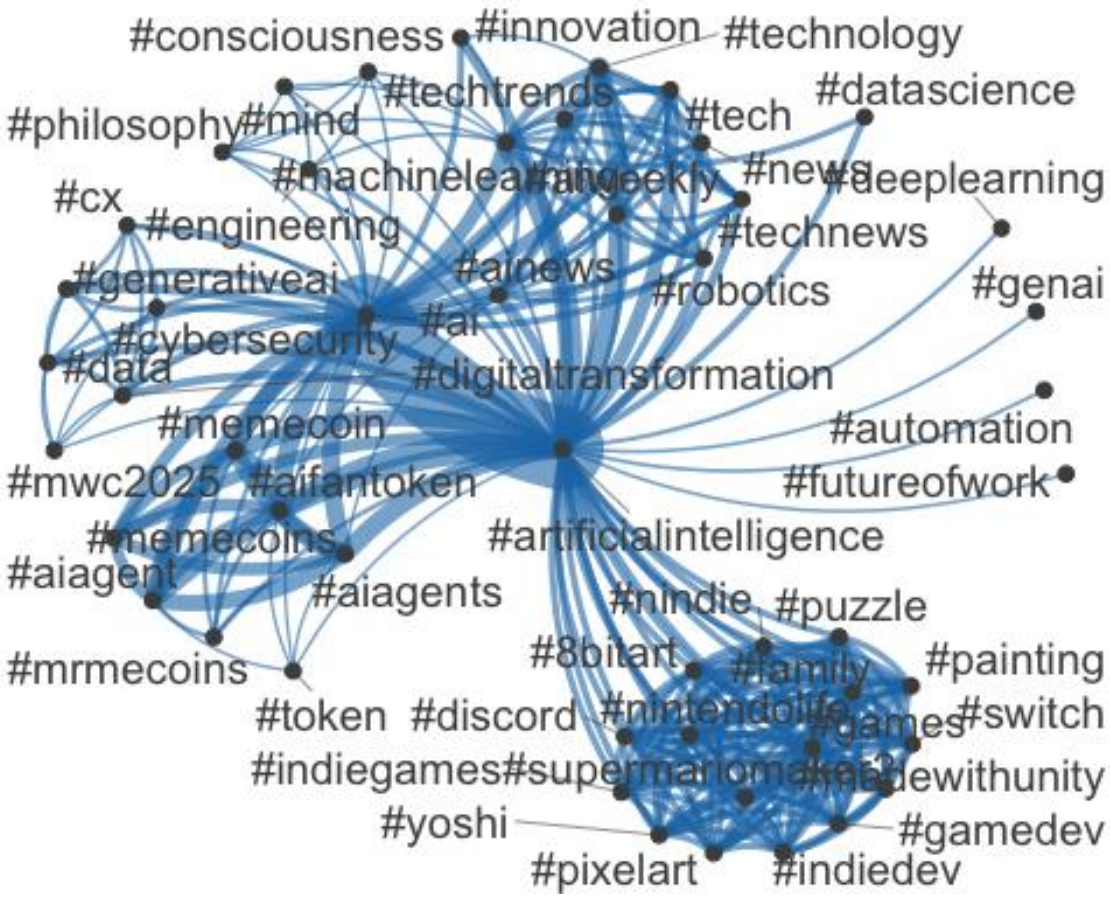
5.3 Rete di Hashtag

L'analisi della rete di hashtag evidenzia le connessioni tra i termini più frequenti nei tweet riguardanti l'intelligenza artificiale. Il nodo centrale è #artificialintelligence, che funge da punto di snodo tra i cluster principali, in particolare ne sono stati evidenziati tre:

- Cluster tecnologico e di business: include hashtag come #ai, #machinelearning, #technews, #digitaltransformation, #robotics, #cybersecurity, che indicano discussioni su innovazione, automazione e impatti sul lavoro e sulla società.

- Cluster legato a criptovalute e trend emergenti: comprende #memecoins, #aifantoken, #mrmemecoins, suggerendo un'intersezione tra IA e il mondo delle criptovalute e dei token digitali.
- Cluster gaming e creatività: include hashtag come #indiegame, #gamedev, #pixelart, #nindie, suggerendo un utilizzo dell'IA in ambiti creativi e videoludici.

```
library(quanteda.textplots)
# Seleziona solo gli hashtag più frequenti (es. quelli che appaiono almeno 5 volte)
tag_dfm <- dfm_trim(tag_dfm, min_termfreq = 5)

# Crea la matrice di co-occorrenza
tag_fcm <- fcm(tag_dfm)

# Filtra per i top hashtag che vuoi visualizzare
top_terms <- names(topfeatures(tag_dfm, 50))  # Mantieni solo i 50 hashtag più frequenti
topgat_fcm <- fcm_select(tag_fcm, pattern = top_terms)

# Plotta la rete con le nuove impostazioni
textplot_network(topgat_fcm, min_freq = 0.3, edge_alpha = 0.6, edge_size = 3)
```



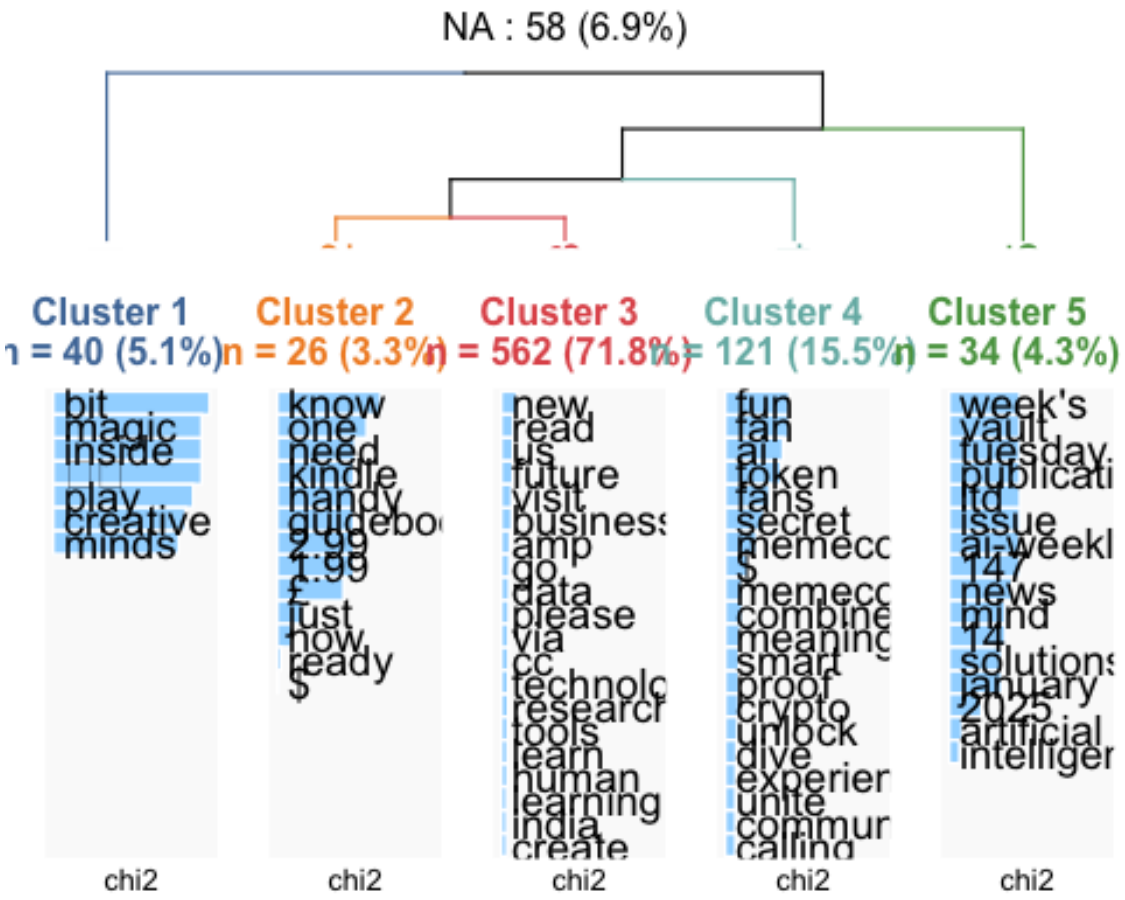
5.4 Reinert Clustering

In questa sezione è stato applicato un algoritmo di clustering (algoritmo di Reinert) per raggrupare tweet simili in insiemi distinti detti cluster. La tipologia di clustering applicato è un clustering gerarchico che costruisce una struttura ad albero (dendogramma) che mostra le relazioni tra i punti dati a livelli diversi. Non è richiesto di specificare il numero di cluster in anticipo. In particolare è un clustering gerarchico di tipo divisivo in cui si parte da un unico cluster contenente tutti i punti e lo si suddivide iterativamente per ottenere cluster più piccoli. È basati sul calcolo del Chi-quadrato, l'obiettivo è associare le unità che hanno valori di Chi-quadrato più alto, cioè quelle maggiormente correlate tra loro.

L'algoritmo di Reinert è un importante approccio gerarchico divisivo che si distingue per la sua applicazione nell'analisi dei dati documentali. La caratteristica distintiva del metodo è che lavora con dati qualitativi che sebbene siano di natura testuale non dovrebbero essere decontestualizzati, poichè questo aumenterebbe l'ambiguità. In altre parole, mentre si crea un cluster si rischia di perdere il contesto che dà significato al dato. Questo principio si ricollega al problema dell'encoding nei modelli di deep learning.

```
library(quanteda)
library(rainette)

##
## Attaching package: 'rainette'

## The following object is masked from 'package:stats':
##
##     cutree

# Creazione del corpus
mycorp <- corpus(tweet_df, text_field = "text")

# Creazione dei tokens per i tweet testuali
tok_text <- tokens(mycorp, remove_punct = TRUE)
tok_text <- tokens_tolower(tok_text)
tok_text <- tokens_remove(tok_text, stopwords("en"))
tok_text <- tokens_remove(tok_text, "#*")  # Rimuove gli hashtag
tok_text <- tokens_remove(tok_text, "@*")  # Rimuove le chiocciole
tok_text <- tokens_remove(tok_text, pattern = "http[s]?://\\S+", valuetype =
"regex")


# Creazione della dfm per i tweet testuali
dtm_text <- dfm(tok_text)
dtm_text <- dfm_trim(dtm_text, min_docfreq = 10)


# Clustering con rainette per i tweet testuali
res_text <- rainette(dtm_text, k = 5)
```

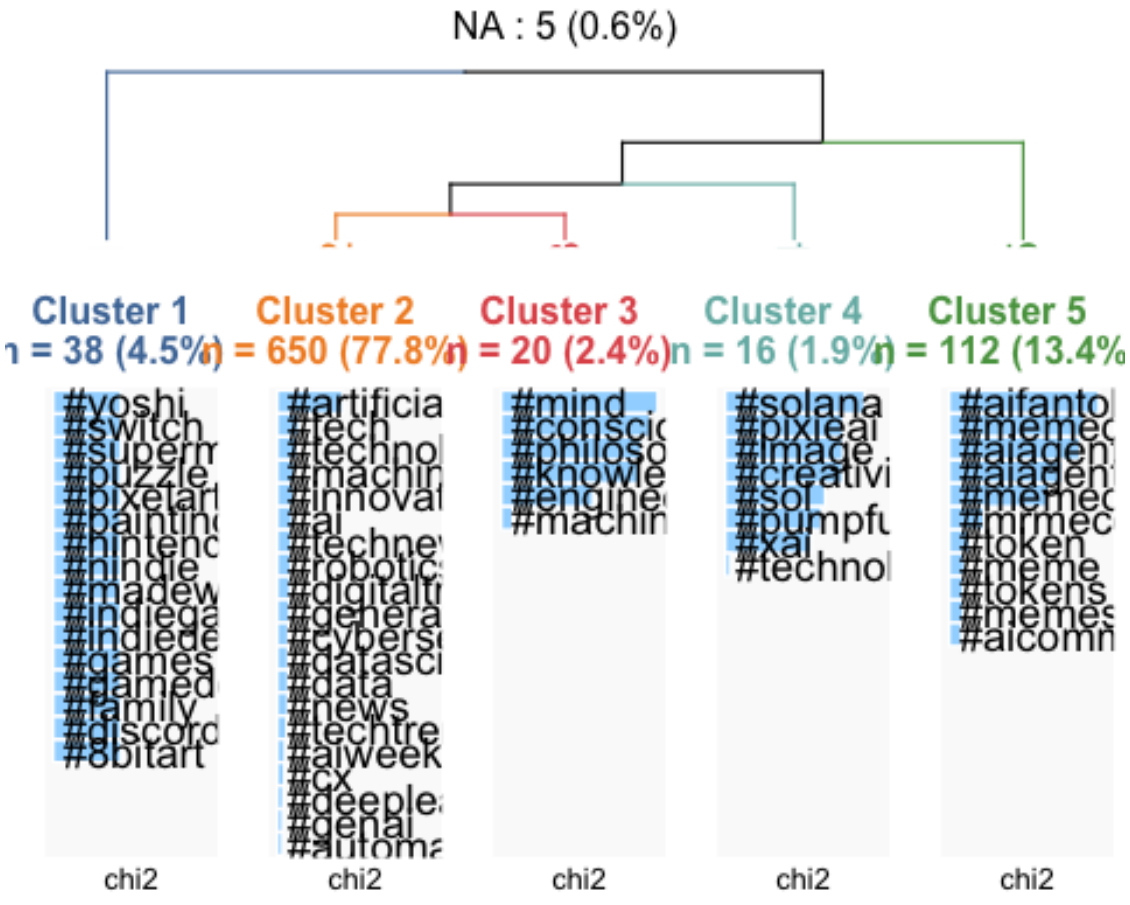
```
## Warning in rainette(dtm_text, k = 5): some documents don't have any term, they
## won't be assigned to any cluster.

##   Clustering...

##   Done.

# Esplorazione e visualizzazione
rainette_explor(res_text, dtm_text, mycorp)

## Loading required package: shiny

##
## Attaching package: 'shiny'

## The following object is masked from 'package:qdapRegex':
##
##     validate

##
## Listening on http://127.0.0.1:4236

##
## -------- Start exported code --------
##
## ## Clustering description plot
## rainette_plot(
##   res_text, dtm_text, k = 4,
##   n_terms = 20,
##   free_scales = FALSE,
##   measure = "chi2",
##   show_negative = FALSE,
##   text_size = 12
## )
## ## Groups
## cutree_rainette(res_text, k = 4)
##
## --------- End exported code ---------

rainette_plot(res_text, dtm_text, k = 5, n_terms = 20, free_scales = FALSE,
measure = "chi2", show_negative = FALSE, text_size = 12)
```




```
# Ottenere i gruppi
groups_text <- cutree_rainette(res_text, k = 5)

# Creazione dei tokens solo per gli hashtag
tok_text_hastag <- tokens(mycorp, remove_punct = TRUE)
tok_text_hastag <- tokens_tolower(tok_text_hastag)
tok_text_hastag <- tokens_remove(tok_text_hastag, stopwords("en"))
tok_hashtag <- tokens_select(tok_text_hastag, pattern = "#*", selection = "keep")

# Creazione della dfm per gli hashtag
dtm_hashtag <- dfm(tok_hashtag)
dtm_hashtag <- dfm_trim(dtm_hashtag, min_docfreq = 10)

# Clustering con rainette sugli hashtag
res_hashtag <- rainette(dtm_hashtag, k = 5)

## Warning in rainette(dtm_hashtag, k = 5): some documents don't have any term,
## they won't be assigned to any cluster.

##   Clustering...

##   Done.

# Esplorazione e visualizzazione
rainette_explor(res_hashtag, dtm_hashtag, mycorp)
```

```
##
## Listening on http://127.0.0.1:4236

##
## -------- Start exported code --------
##
## ## Clustering description plot
## rainette_plot(
##   res_hashtag, dtm_hashtag, k = 5,
##   n_terms = 20,
##   free_scales = FALSE,
##   measure = "chi2",
##   show_negative = FALSE,
##   text_size = 12
## )
## ## Groups
## cutree_rainette(res_hashtag, k = 5)
##
## --------- End exported code ---------
```

```
rainette_plot(res_hashtag, dtm_hashtag, k = 5, n_terms = 20, free_scales = FALSE,
measure = "chi2", show_negative = FALSE, text_size = 12)
```




```
# Ottenere i gruppi
groups_hashtag <- cutree_rainette(res_hashtag, k = 5)
```

L’analisi di Reinert è stata applicata ai testi dei tweet e, separatamente, agli hashtag contenuti nei tweet, tutti incentrati sul tema dell’intelligenza artificiale. Questo ha permesso di individuare i principali cluster semantici emergenti.

Nel clustering basato sui testi, sono stati identificati quattro gruppi:

Cluster 1: raccoglie parole legate alla creatività e all’immaginazione (magic, play, creative), suggerendo un interesse per l’aspetto innovativo e ispirazionale dell’IA.

Cluster 2: il più numeroso, contiene termini generici legati a tecnologia, futuro e business (future, business, data, technology), indicando un focus su applicazioni pratiche e sviluppo del settore.

Cluster 3: include parole connesse a criptovalute e intelligenza artificiale (token, memecoin, crypto, unlock), segnalando un’intersezione tra AI e finanza decentralizzata.

Cluster 4: presenta riferimenti temporali e editoriali (week’s, publication, AI-weekly, January 2025), suggerendo una dimensione informativa e divulgativa sull’IA.

Nel clustering basato sugli hashtag, sono emersi cinque gruppi distinti:

Cluster 1: hashtag legati al mondo del gaming e della pixel art (#pixelart, #games, #indiedev), evidenziando l’uso dell’IA in ambito creativo e videoludico.

Cluster 2: il più ampio, comprende termini associati a tecnologia, innovazione e ricerca sull’AI (#artificial, #machine, #robotics, #datascience).

Cluster 3: raccoglie concetti legati alla mente e alla filosofia dell’IA (#mind, #conscious, #philosophy), suggerendo un interesse per gli aspetti cognitivi e teorici.

Cluster 4: combina elementi creativi e blockchain (#pixelart, #image, #solana), indicando interazioni tra AI, arte e tecnologie decentralizzate.

Cluster 5: si concentra su criptovalute e meme coin (#meme, #tokens, #memecommunity), sottolineando il legame tra AI e cultura finanziaria digitale.

L’analisi evidenzia come i testi dei tweet mostrino una varietà tematica più ampia, mentre gli hashtag si concentrano su specifiche aree di applicazione dell’intelligenza artificiale, tra cui gaming, filosofia, ricerca tecnologica e criptovalute.

CONCLUSIONI

L’analisi condotta ha permesso di esplorare l’uso dell’intelligenza artificiale nelle aziende europee, evidenziando le principali tendenze e ambiti di applicazione. Successivamente, attraverso l’analisi testuale dei documenti, sono stati individuati i temi ricorrenti e le prospettive emergenti. Infine, l’analisi del sentiment e il clustering dei tweet hanno confermato il crescente interesse per l’AI, con una discussione ampia e diversificata che spazia dalla ricerca tecnologica alle implicazioni economiche e sociali.

L’intelligenza artificiale si conferma un tema di grande attualità e in continua evoluzione. Il suo utilizzo nelle aziende è destinato a crescere, portando innovazioni e nuove sfide. Sarà interessante osservare, nei prossimi anni, come le percezioni e le applicazioni dell’AI si trasformeranno, sia nel dibattito pubblico che nelle strategie aziendali.