Web Analytics e Analisi Testuale 2023/2024

Mariella Ruiu (11/82/00362)

Sentiment Analysis, Emotion Analysis e Topic Modeling del podcast One More Time di Luca Casadei: dall'audio testo dei video ai commenti degli utenti su YouTube

INTRODUZIONE

Descrizione del Podcast:

- Interviste con personaggi noti che condividono esperienze di vita significative
- Più di 300.000 iscritti su YouTube, milioni di visualizzazioni alle spalle

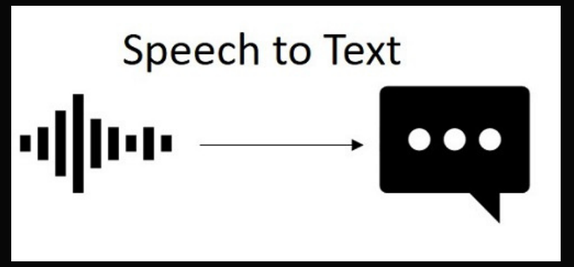
Obiettivi del Progetto (diviso in due parti):

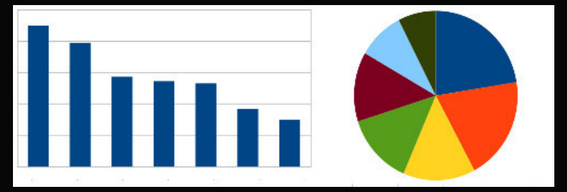
- Analizzare il testo trascritto dei video e i commenti degli utenti su YouTube:
  - Analisi del sentiment e delle emozioni
  - Analisi dei topic
  - Analisi comparativata tra le interviste

Prima Parte del Progetto: Analisi del Testo Trascritto sugli Audio dei Video

Struttura:

- *Selezione di due video popolari*
- *Estrazione e trascrizione dell'audio in testo*
- *Sentiment Analysis e Emotion Analysis per frase*
- *Analisi temporale del sentiment e delle emozioni*
- *Topic Modeling per identificare i temi chiave*
- *Confronto tra i video analizzati*





Conversione audio (formato .mp3) in testo (formato .txt)

Ambiente di sviluppo:

- Google Colaboratory

```
!pip install git+https://github.com/openai/whisper.git
```

Modello di apprendimento automatico:

- Whisper AI - model large

```
!sudo apt update && sudo apt install ffmpeg
```

Acceleratore hardware (Runtime)

- T4 GPU

```
!whisper "/content/Bianca Balti da squatter alle passerelle più prestigiose del mondo - One More Time.mp3" --model large
```

BiancaBalti_model_large

StevenBasalari_model_large

"ONE MORE TIME"

Analisi del sentiment e delle emozioni

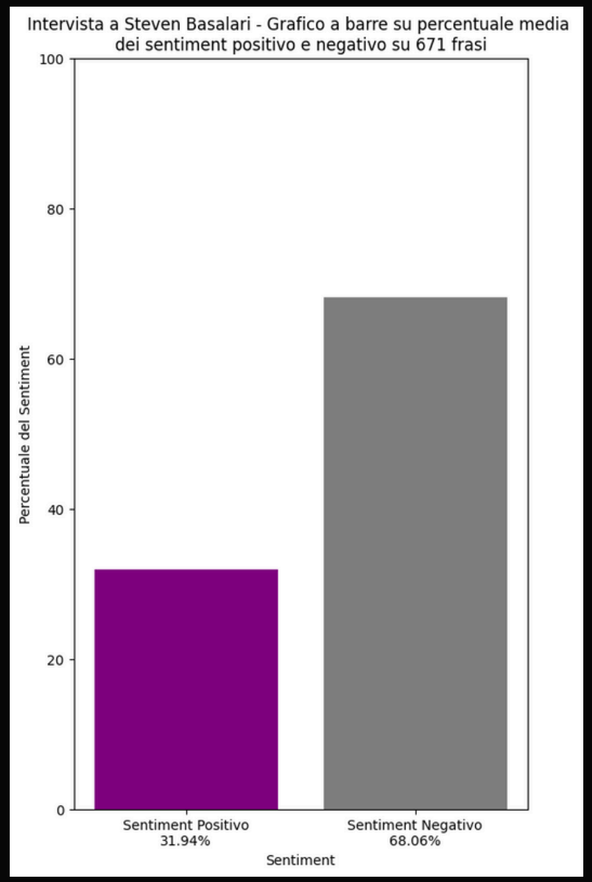
Modello pre-addestrato per la sentiment analysis di Hugging Face

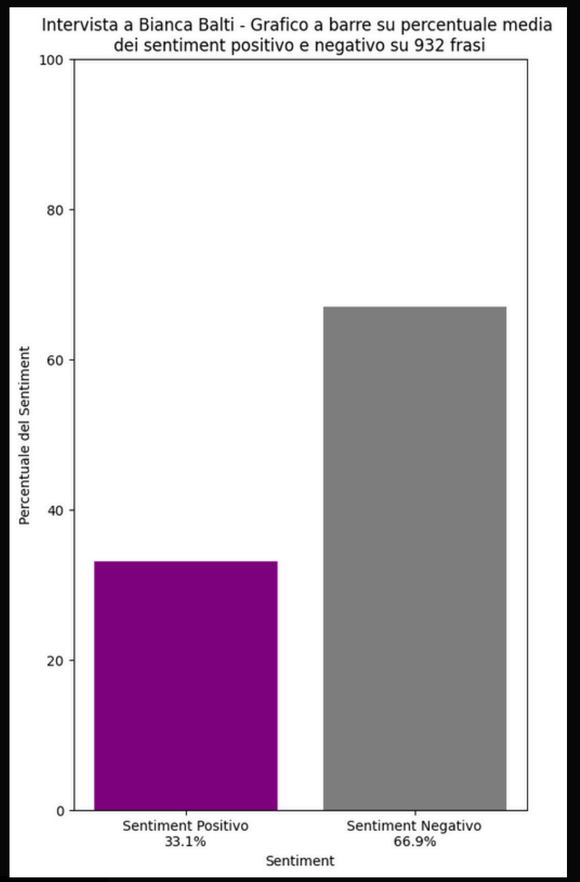
- MilaNLProc/feel-it-italian-sentiment
- Restituisce due valori complementari numerici (positivo, negativo)

Modello pre-addestrato per l'emotion analysis dalla libreria feel_it

- EmotionClassifier
- Restituisce un tipo di emozione (gioia, tristezza, rabbia, paura)

Visualizzazioni grafiche: analisi dei sentiment (Positivo, Negativo)





"ONE MORE TIME"

Analisi comparativa: t-test per confrontare i sentiment negativi delle due interviste

```
Statistiche del t-test per il sentiment negativo:
Statistica t: -0.49658481938080684
Valore p_value: 0.61955687945389 75
Non c'è una differenza significativa nei sentiment negativi tra le due interviste.
```

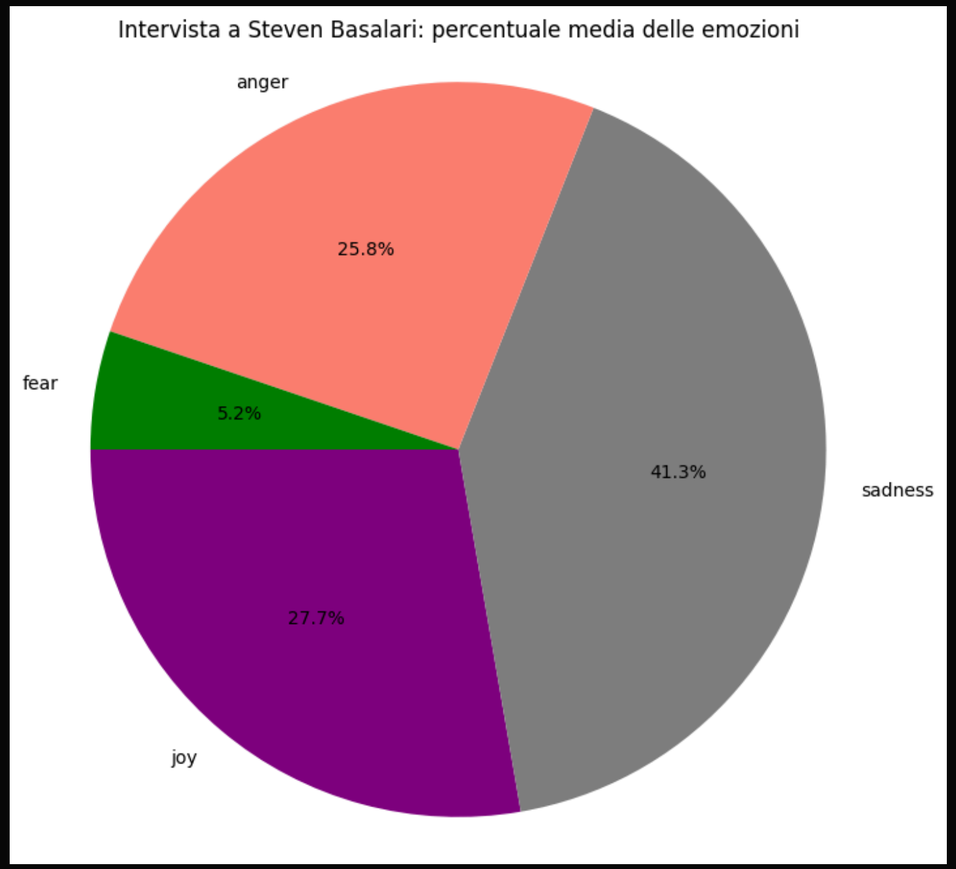
Statistica t:

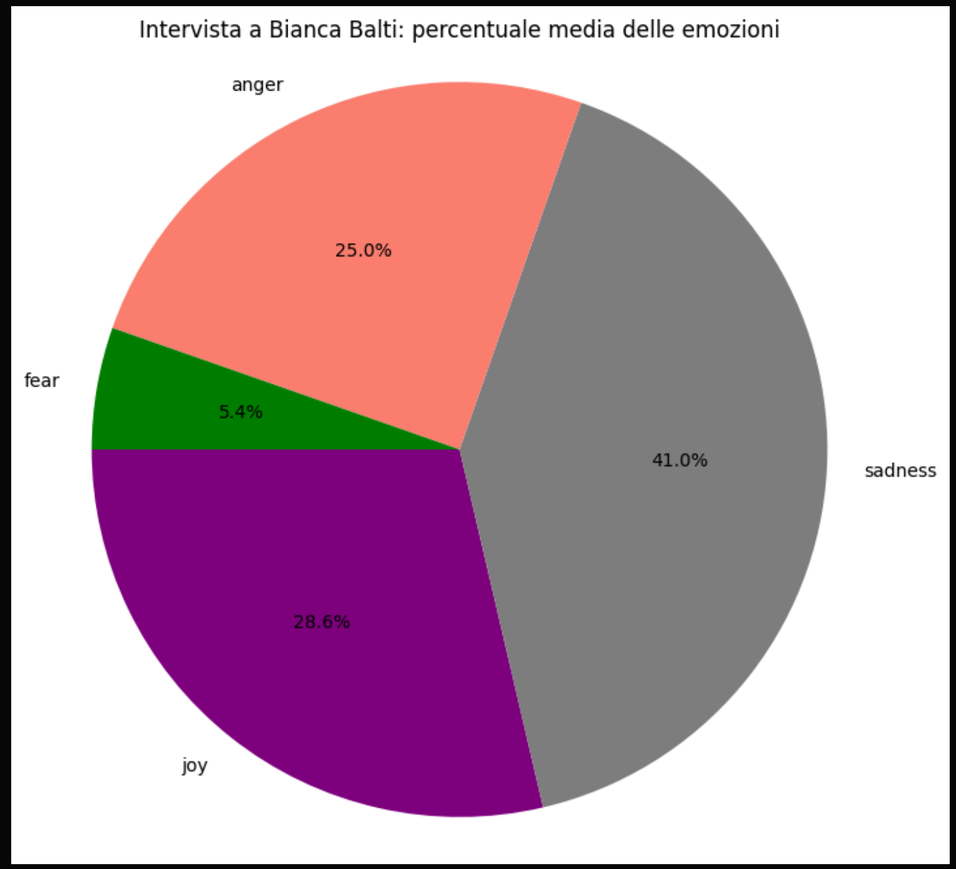
- indica che le medie dei sentiment negativi delle due interviste sono simili

P-value:

- p-value > α (0.05) rigetta l’ipotesi nulla
- Non c'è una differenza significativa nei sentiment negativi tra le due interviste quindi l’ipotesi nulla viene rigettata

Visualizzazioni grafiche: analisi delle emozioni (gioia, tristezza, paura, rabbia)





Analisi comparativa: test chi-quadro per confrontare le distribuzioni delle emozioni predette delle interviste

```
Risultati del test del chi-quadro per confronto delle emozioni predette:
Chi-quadro: 0.237749070988434 67
Valore p_value: 0.9712768374590375
Non c'è una differenza significativa nelle distribuzioni delle emozioni predette tra le due interviste.
```

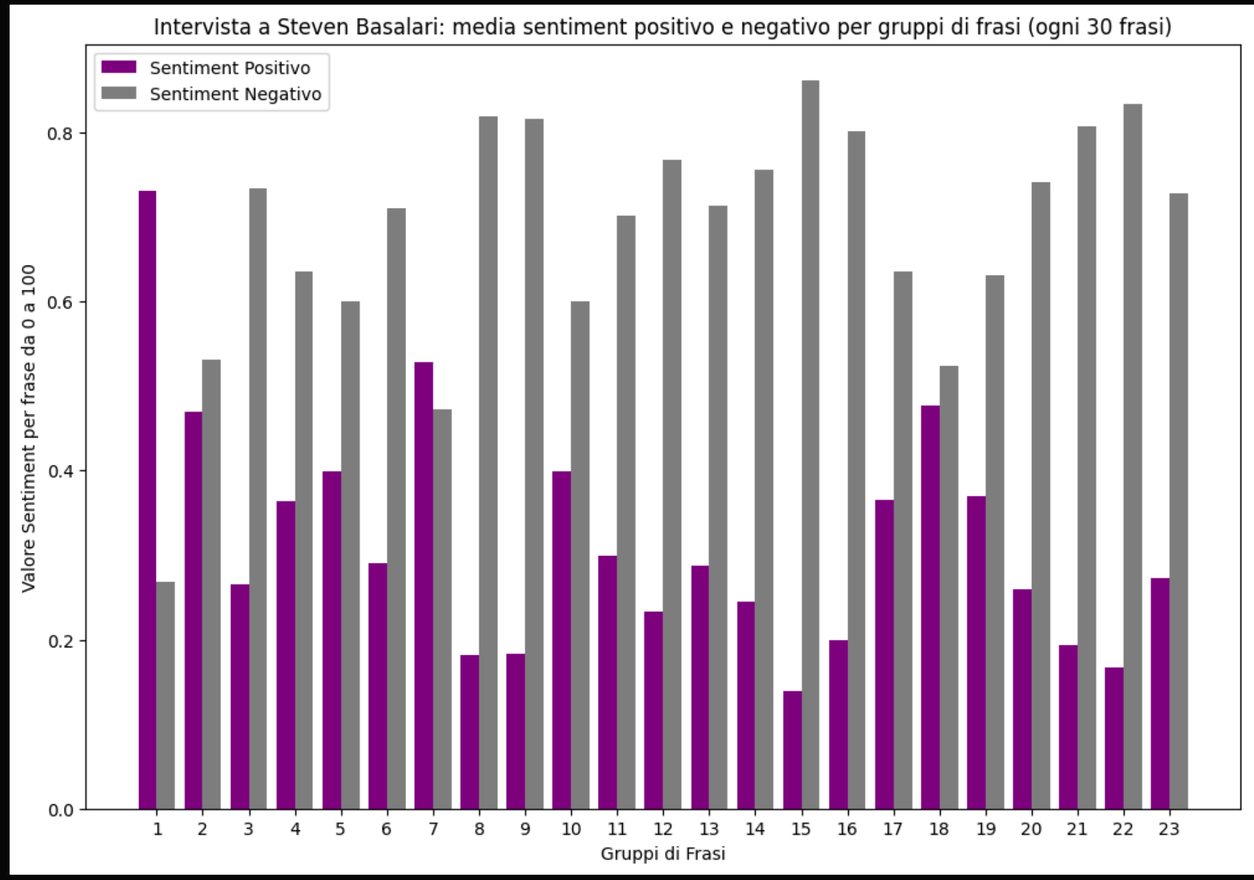
Valore Chi-quadro:

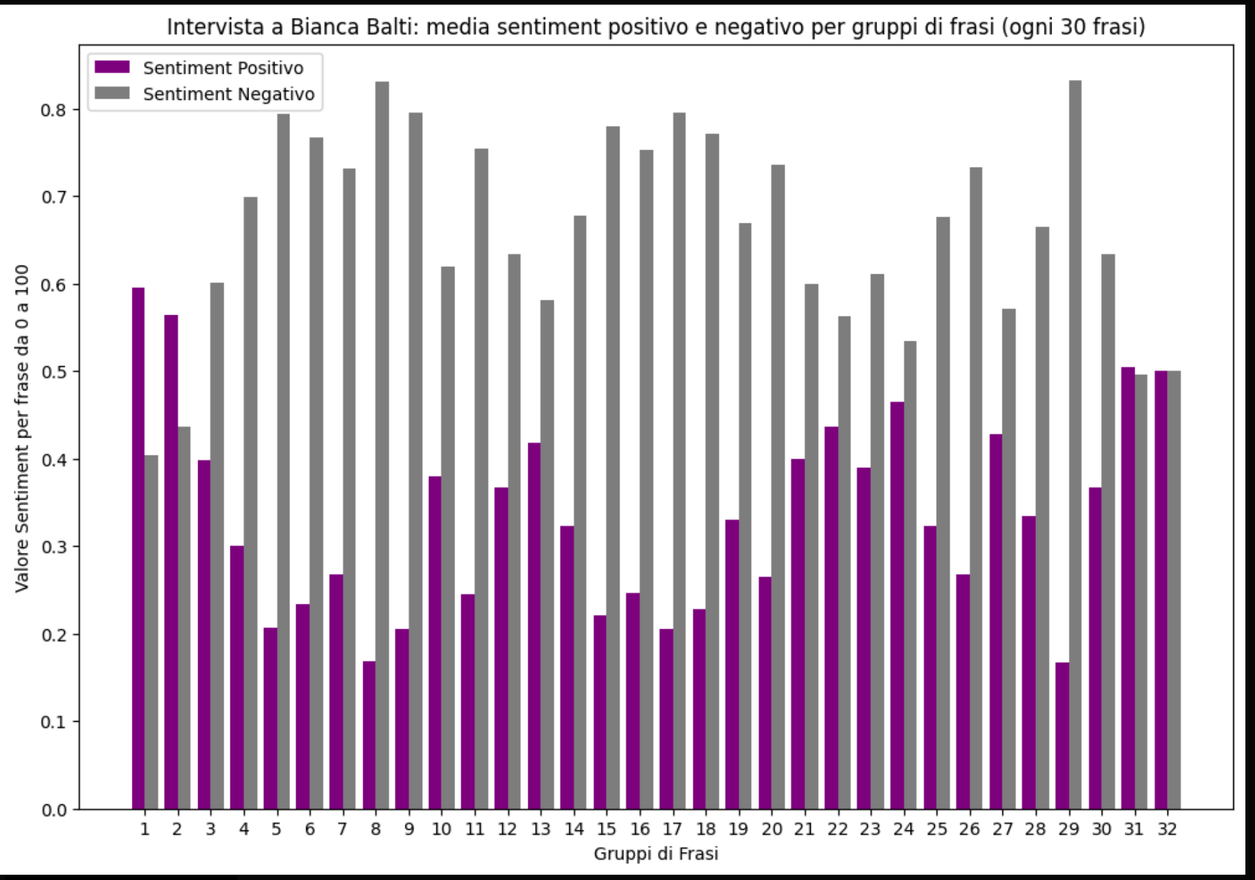
- valore basso, indica che le distribuzioni osservate delle emozioni predette nelle due interviste sono molto simili.

P-value:

- p-value > α (0.05) rigetta l'ipotesi nulla
- non c'è una differenza significativa tra le distribuzioni delle emozioni predette nelle due interviste e quindi l'ipotesi nulla viene rigettata.

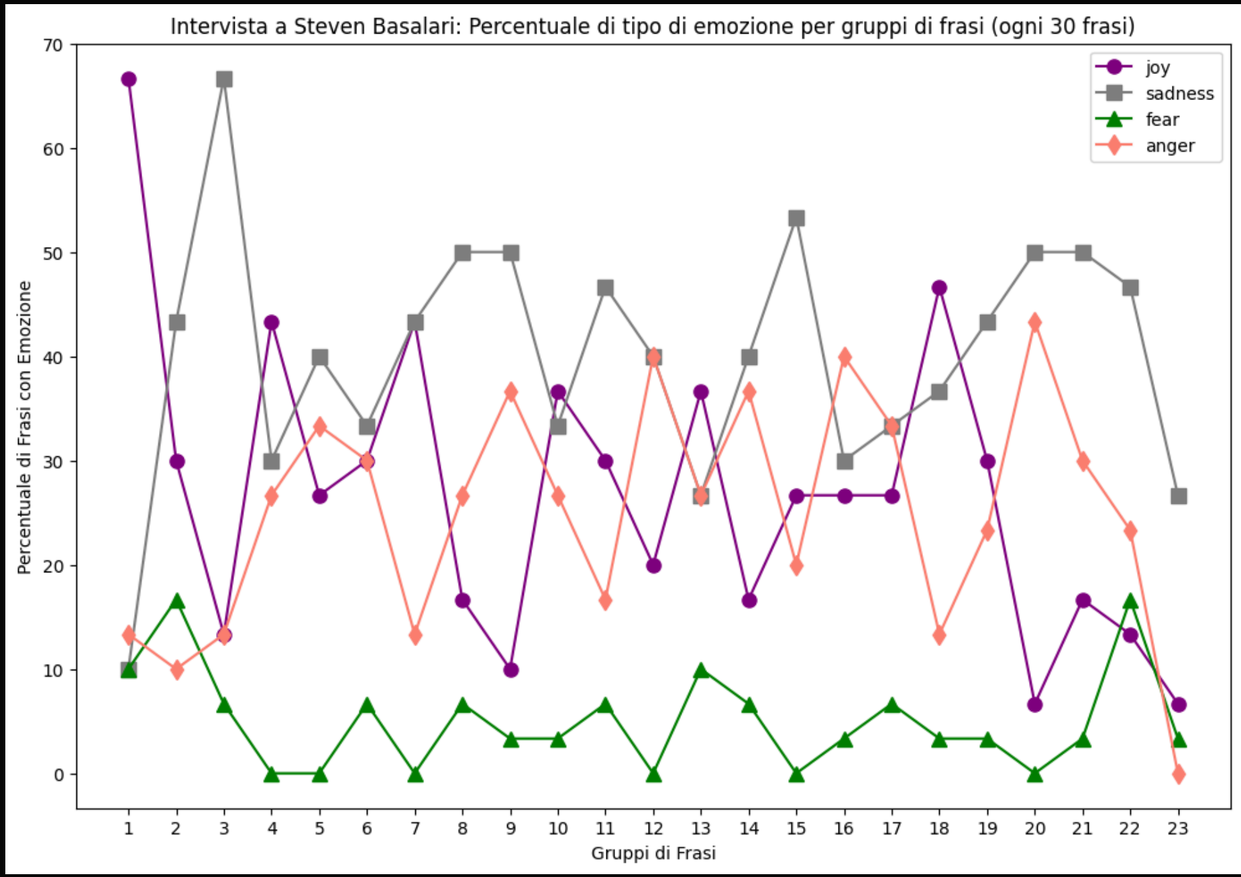
Visualizzazioni grafiche: distribuzione dei sentiment durante l’intervista

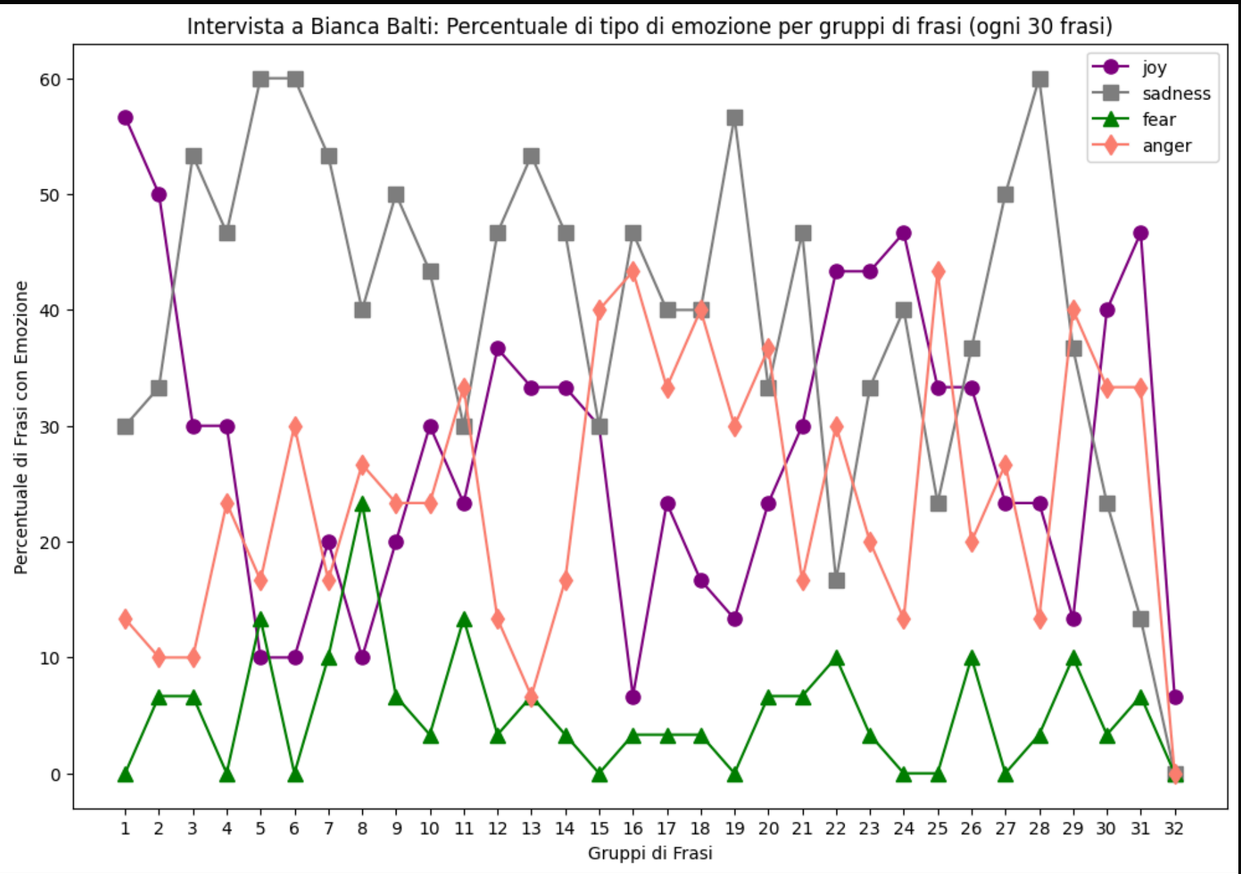




“ONE MORE TIME”

Visualizzazioni grafiche: distribuzione delle emozioni durante l’intervista



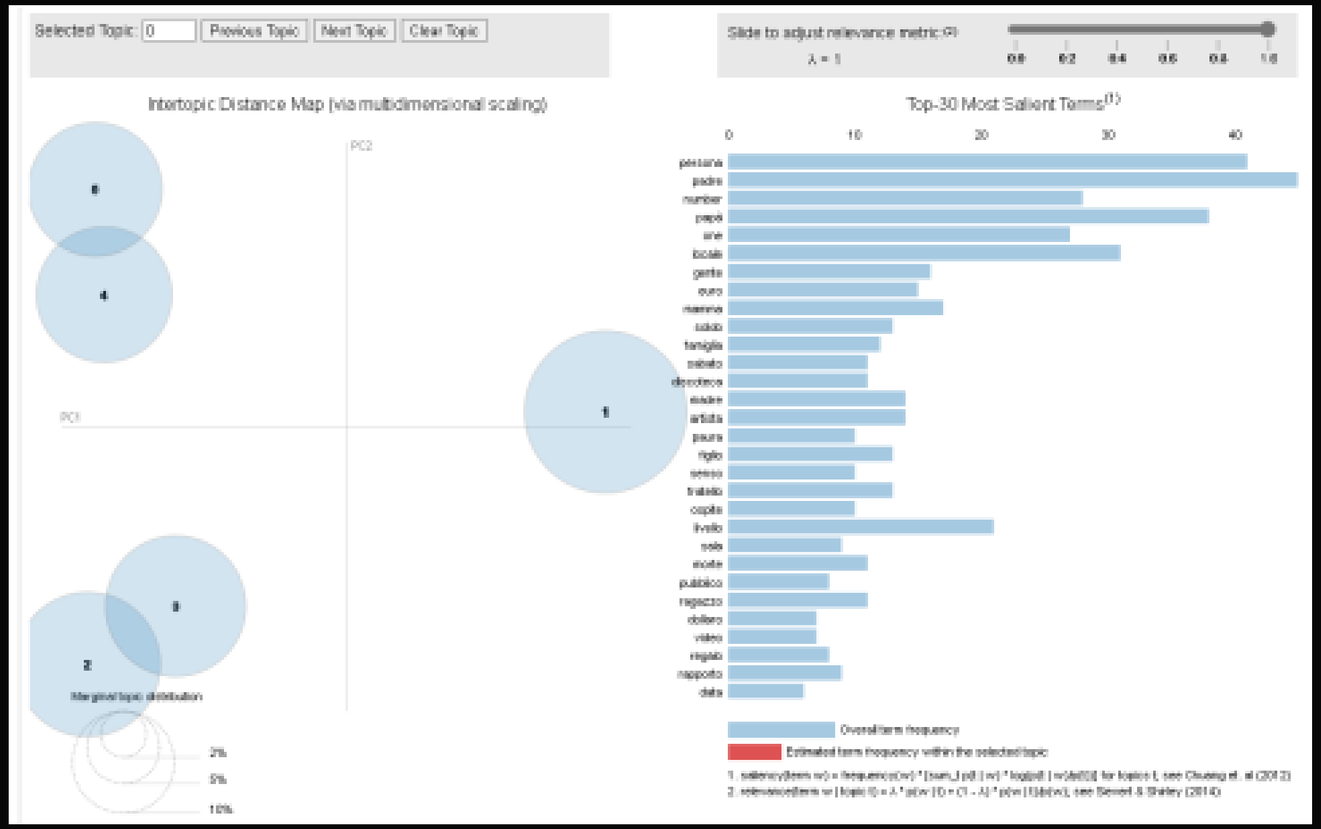


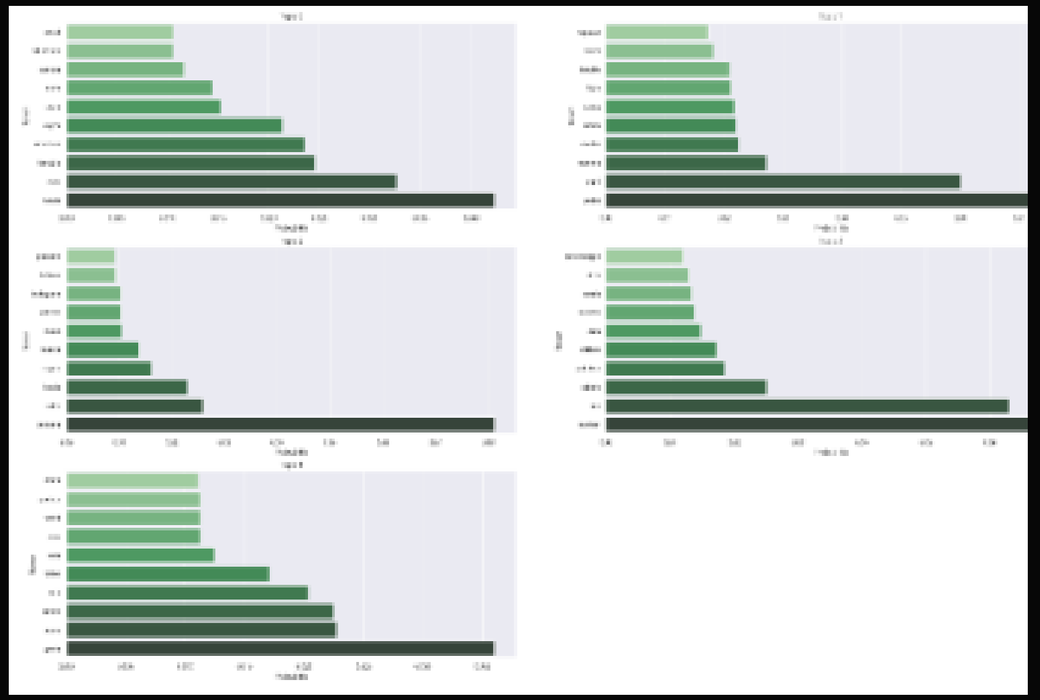
“ONE MORE TIME”

Topic modeling sui testi delle interviste

- Modello LDA (Latent Dirichlet Allocation) di Gensim
- estrazione di 5 topic per ogni intervista
- creazioni visualizzazioni di topic per ogni intervista:
  - *interfaccia interattiva dei topic con pyLDAvis*
  - *grafici a barre per i topic estratti*
  - *word cloud per i topic*
- livello di coerenza dei topic estratti da LDA per ogni intervista.

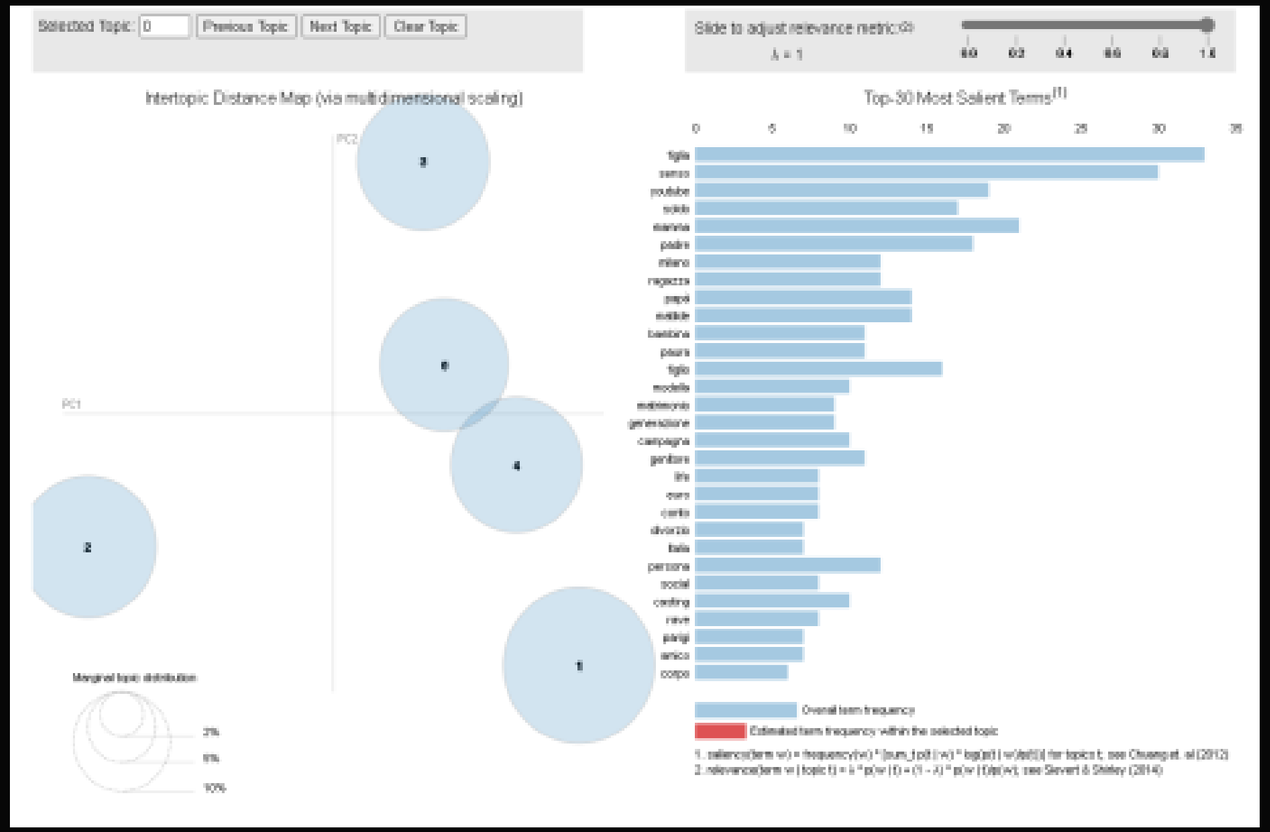
Risultati immagini dei metodi appena elencati dell’intervista di Steven Basalari

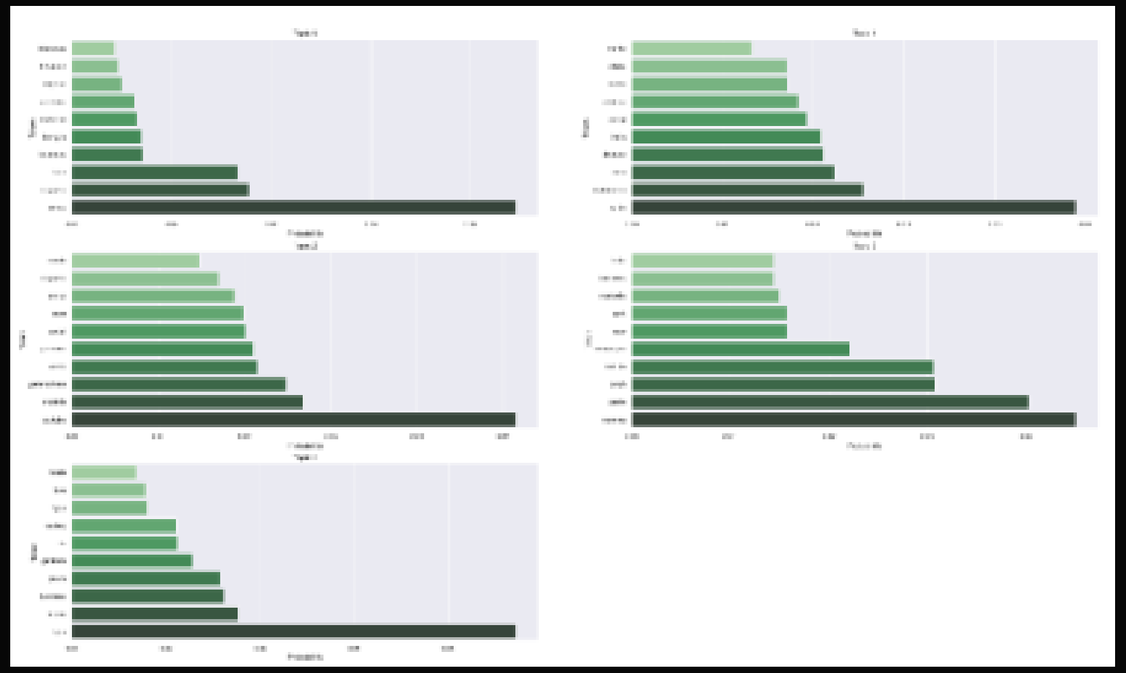




“ONE MORE TIME”

Risultati immagini dei metodi appena elencati dell’intervista di Bianca Balti





Analisi comparativa: test Chi-quadro per confrontare i topic dominanti delle interviste

```
Analisi comparativa dei topic dominanti:
Statistica del test del Chi-quadrato: 36.47316053038837
P-value: 2.3124997315307777e-07
C'è una relazione significativa tra i testi e i topic dominanti.
```

Valore Chi-quadro:

- mostra quanto i topic dominanti nei testi di due modelli differiscano da quanto ci si aspetterebbe senza una relazione tra i testi e i topic

P-value:

- p-value < α (0.05) rigetta l'ipotesi nulla
- c'è una relazione significativa tra i testi e i topic dominanti, quindi i dati suggeriscono che i topic estratti sono significativamente associati ai testi analizzati.

Livello di coerenza dei topic estratti per ogni intervista

- Livello coerenza Steven Basalari: 0.38835003530422 95
- Livello coerenza Bianca Balti: 0.4253803064229412

Conclusioni sulla prima parte del progetto

Difficoltà e fonti di inattendibilità:

- Trascrizione automatica di audio con espressioni in dialetto e lingue straniere.
- Errori del modello speech-to-text (Whisper), che introducono imprecisioni e rumore nei testi.

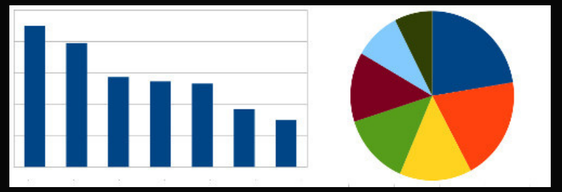
Limitazioni dei modelli di analisi:

- Modelli addestrati di sentiment analysis, emotion analysis e topic modeling potrebbero non funzionare bene su testi rumorosi e non perfettamente puliti.

Seconda Parte del Progetto: Analisi dei Commenti degli Utenti su YouTube

Struttura:

- *Selezione dei dieci video più commentati.*
- *Estrazione dei commenti dai video.*
- *Sentiment Analysis e Emotion Analysis per commento.*
- *Topic Modeling per identificare i temi chiave.*
- *Confronto tra i commenti dei video analizzati.*



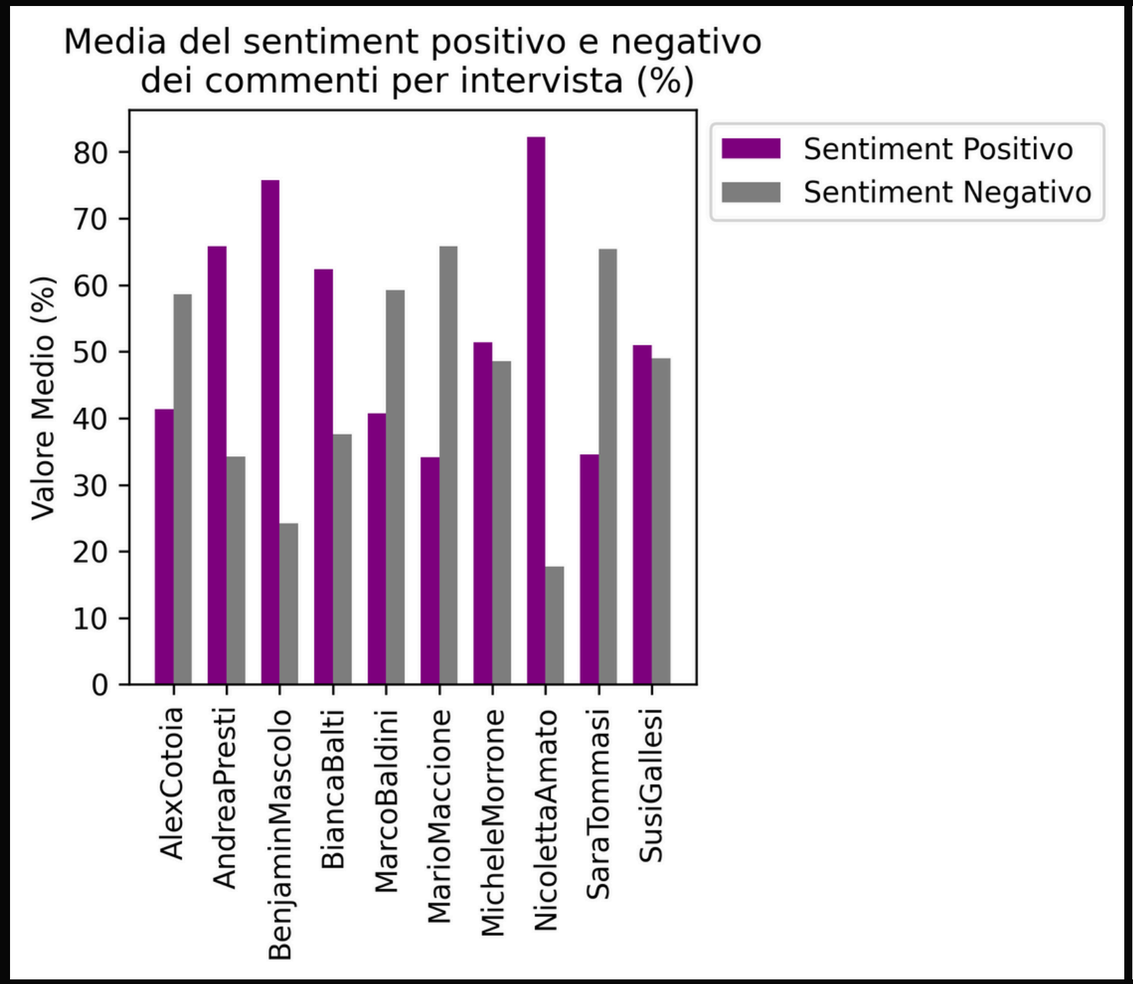
Estrazione dei commenti segli utenti da video su YouTube

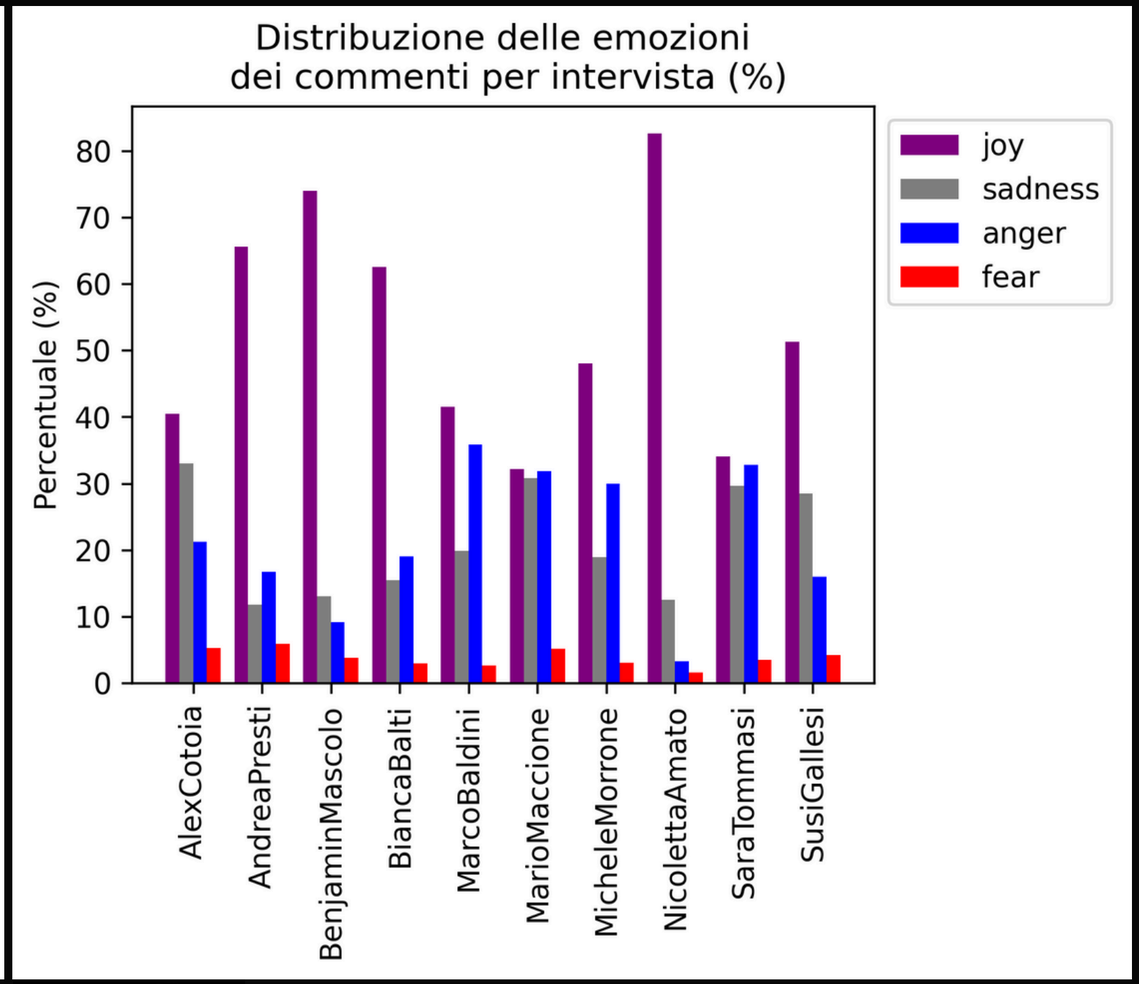
- *googleapiclient*
- *Creazione di un file CSV per ogni intervista che conterrà per ogni commento:*
  - *nome utente, testo del suo commento, data di pubblicazione del commento*

Pulizia commenti

- *emoji*
- *tag HTML*
- *sequenze di caratteri HTML*
- *commenti del canale*
- *commenti vuoti*
- *commenti corti*

Visualizzazioni grafiche: analisi dei sentiment (Positivo, Negativo) e analisi della distribuzione delle emozioni (gioia, tristezza, rabbia, paura)





"ONE MORE TIME"

Analisi comparativa: test ANOVA per confrontare i sentiment negativi delle interviste

```
               sum_sq      df          F         PR(>F)
C(File)    179.294881     9.0  92.512953  1.802564e-162
Residual  1336.609628  6207.0        NaN            NaN

Interpretazione del risultato dell'ANOVA sui sentiment negativi delle interviste:
Esiste una differenza significativa nei sentiment negativi tra i commenti delle interviste prese in considerazione.
```

PR(>F):

- PR(>F) < α (0.05) rigetta l’ipotesi nulla
- c’è una differenza significativa nei sentiment negativi tra i commenti delle diverse interviste considerate, quindi le differenze osservate non sono casuali, ma probabilmente riflettono vere differenze tra le interviste

Analisi comparativa: test Chi-quadro per analizzare l'associazione tra le emozioni (gioia, tristezza, rabbia, paura) e il sentiment (positivo/negativo) all'interno dei commenti

```
Test del Chi-quadro:
Statistiche del test: 4306.51450693535
Valore p: 0.0
Ci sono evidenze di un'associazione significativa tra il tipo di emozione e il tipo di commento (positivo o negativo).
```

Valore Chi-quadro:

- essendo un valore molto grande indica che c'è una forte connessione tra le emozioni espresse nei commenti e se quei commenti sono positivi o negativi.

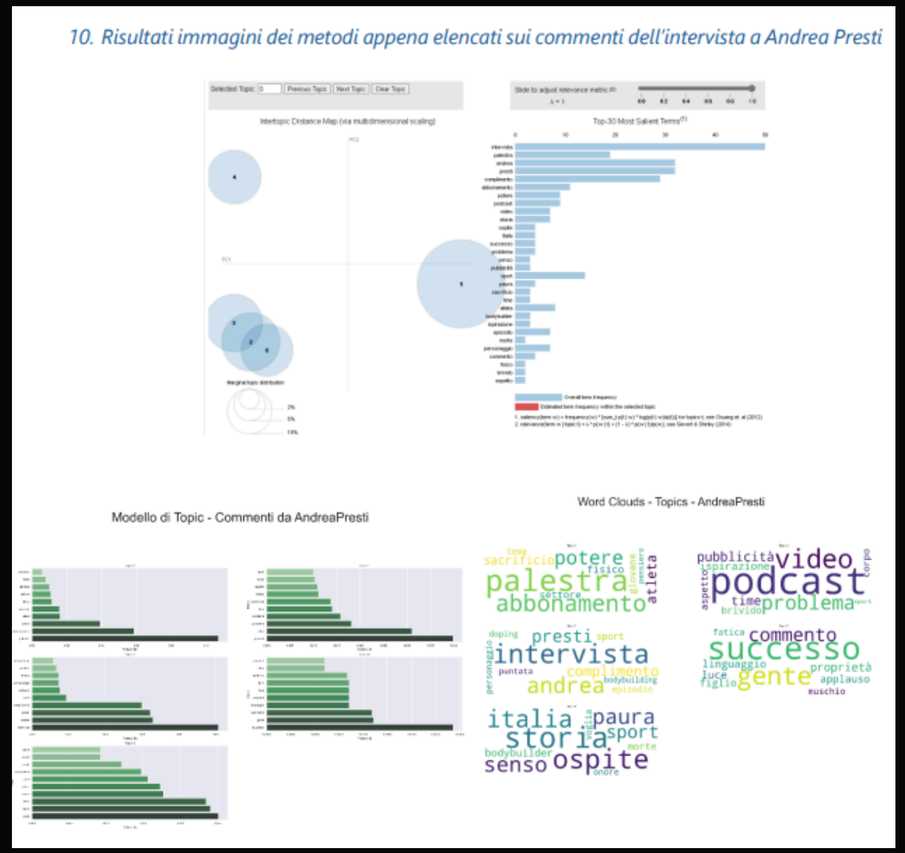
P-value:

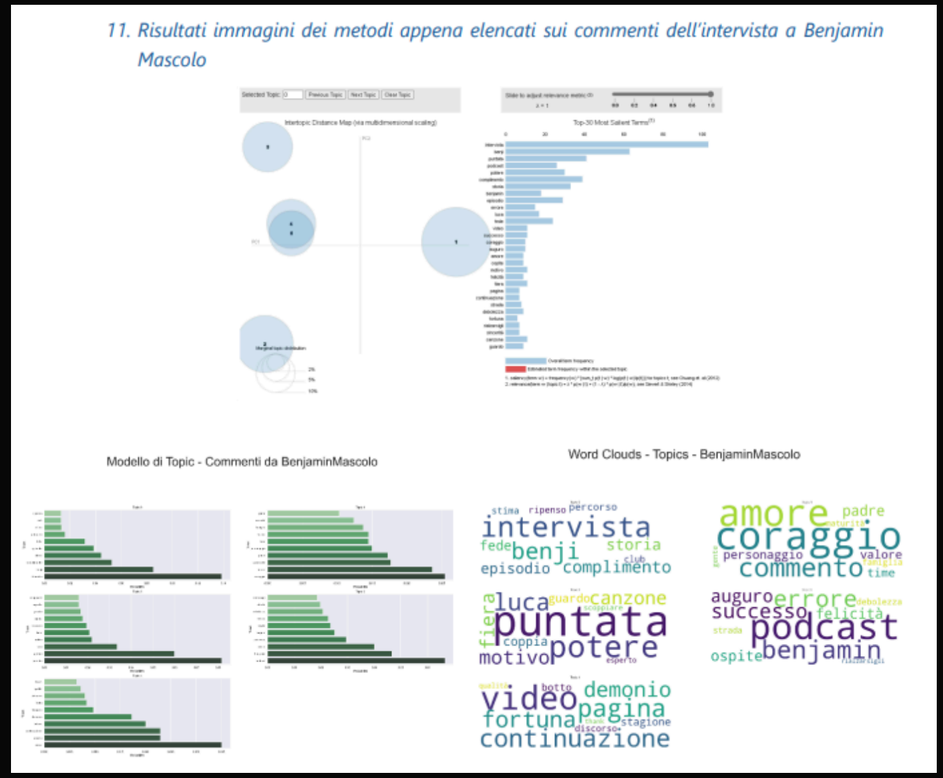
- p-value < α (0.05) rigetta l’ipotesi nulla
- ci sono evidenze di un'associazione significativa tra il tipo di emozione e il tipo di commento

Analisi dei Topic sui Commenti per Ogni Intervista Selezionata

- *Creazione di un'interfaccia interattiva, grafici a barre accostate e word clouds.*

Esempi:

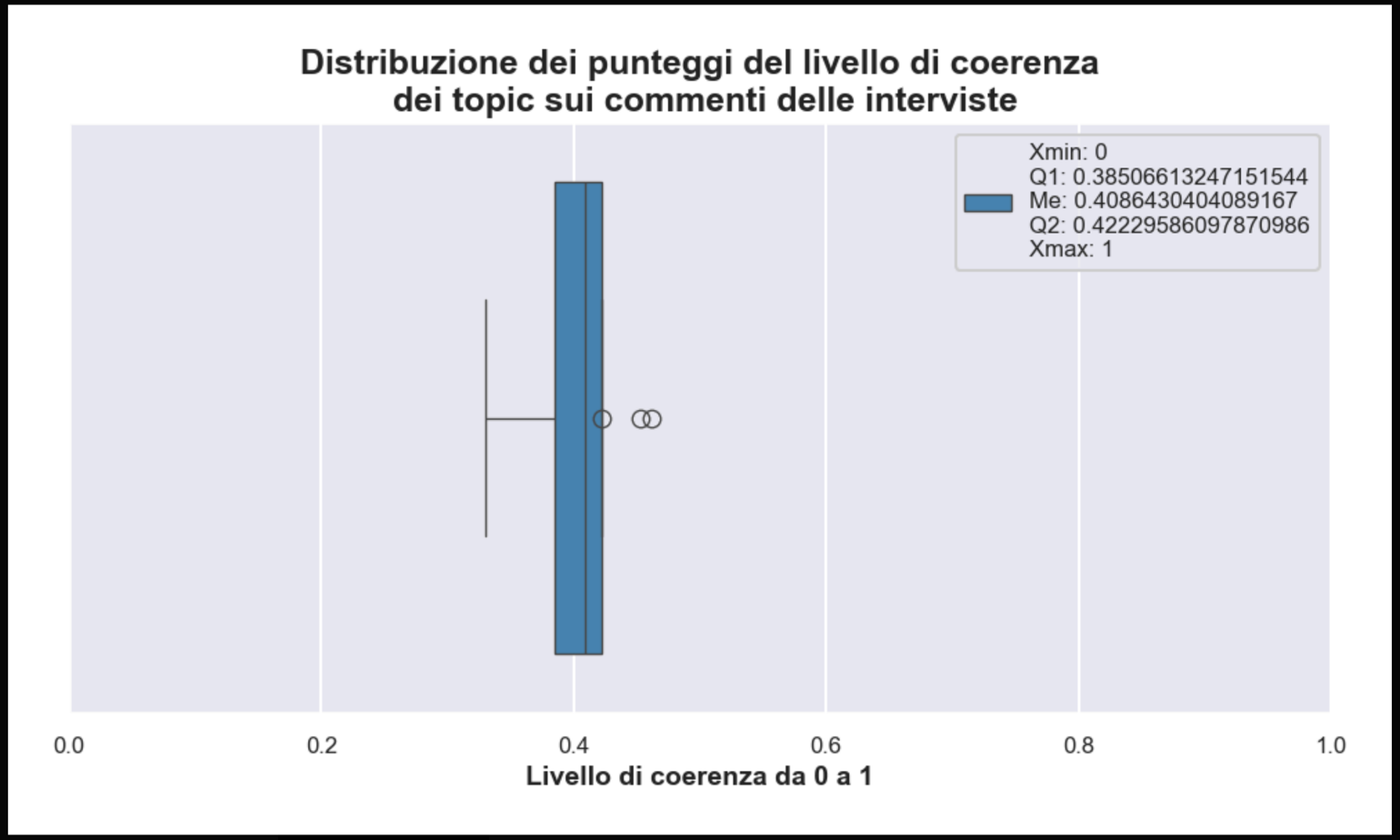




Livello di coerenza dei topic per ogni intervista

```
Punteggi di coerenza in ordine decrescente:
BiancaBalti: 0.4617551383109312
AlexCotoia: 0.4528570754143681
BenjaminMascolo: 0.422401358354336
AndreaPresti: 0.4219793688518315
MicheleMorrone: 0.4132554594958053
MarcoBaldini: 0.4040306213220281
SaraTommasi: 0.3966859152910255
NicolettaAmato: 0.3811928715316788
MarioMaccione: 0.3512003213280078
SusiGallesi: 0.3298651021602524
```

```
Media: 0.40352232320602643
Mediana: 0.4086430404089167
25° percentile: 0.38506613247151544
75° percentile: 0.42229586097870986
Varianza: 0.001532537031957979
Scarto quadratico medio: 0.03914763124325633
```



"ONE MORE TIME"

Conclusioni sulla seconda parte del progetto

Difficoltà e fonti di inattendibilità:

- La lingua italiana dei commenti è contaminata da dialetti e lingue straniere.
- Presenza di errori di battitura e abbreviazioni che compromettono l'accuratezza dei modelli di analisi.

Limitazioni dei modelli di analisi:

- Modelli di sentiment analysis, emotion analysis e topic modeling, addestrati su testi "puliti", potrebbero non funzionare bene su commenti "rumorosi".


Fine
ONE MORE TIME