

# Leaders tweets analysis during Covid-19 pandemic



Simone Contini, Salvatore Calderaro  
Università degli Studi di Palermo



# Introduzione

Considerati i Tweet dei capi di stato dei paesi Italia, Inghilterra, Francia, Brasile e India, in un periodo compreso tra Gennaio 2020 ad oggi, sono state effettuate delle analisi per rilevare i topic di interesse per ciascuno dei quattro periodi analizzati. Inoltre, per ciascuno di questi periodi, sono state effettuate delle emotion analysis dai testi dei tweet.



# Step

1. Scraping dei tweet
2. Cleaning dei dati
3. Rimozione stopwords e lemmatizzazione
4. Estrazione dei topic
5. Creazione di un modello di machine learning per l'emotion detection
6. Emotion analysis dei tweet
7. Plot dei risultati



# Scraping dei tweet

Per mezzo della libreria Twint sono stati estratti tutti i tweet, per un periodo compreso tra Gennaio 2020 e Giugno 2021, dei capi di stato dei seguenti paesi:

- Italia
- Inghilterra
- Francia
- Brasile
- India

Per ciascuno di questi paesi è stato creato un dataset, di cui si è fatta un'ulteriore divisione nei seguenti periodi:

- Primo periodo: 1 Gennaio 2020 - 31 Maggio 2020
- Secondo periodo: 1 Giugno 2020 - 30 Settembre 2020
- Terzo Periodo: 1 Ottobre 2020 - 31 Gennaio 2021
- Quarto Periodo: 1 Febbraio 2021 - 9 Giugno 2021



# Cleaning e preprocessing dei dati

Per ciascuno dei dati ottenuti, è stata effettuata:

- Rimozione dei caratteri di new line
- Rimozione della punteggiatura
- Rimozione dei link
- Rimozione delle emoticon
- Conversione del testo in minuscolo



# Rimozione stopword e lemmatizzazione

Effettuata la tokenizzazione:

- Per la rimozione delle stopwords è stata utilizzata la libreria *NLTK*
- Per la lemmatizzazione è stata utilizzata la libreria *spaCy*

Per entrambe le operazioni, sono stati effettuati opportuni settaggi in base alla lingua del testo (italiano, inglese, francese, portoghese).

# Vettorizzazione dei testi

Consiste nel trasformare un testo in un vettore numerico.

Dato un corpus formato da  $N$  documenti e un vocabolario costituito da  $M$  termini, il risultato sarà una matrice sparsa, di dimensione  $N \times M$ , dove la riga  $i$ -esima corrisponde alla rappresentazione vettoriale del documento  $i$ -esimo.

```
diz = array( [  
    'Oggi è una bella giornata.',  
    'Oggi non è una bella giornata.',  
    'Oggi sta piovendo.',  
    ]  
)
```

['bella', 'giornata', 'non', 'oggi', 'piovendo', 'sta', 'una']

[1 1 0 1 0 0 1]

[1 1 1 1 0 0 1]

[0 0 0 1 1 1 0]



# Latent Dirichlet Allocation (LDA)

LDA è un modello probabilistico generativo per collezioni di dataset discreti come ad esempio corpus di documenti. Viene anche adoperato per l'estrazione di topic astratti da una collezione di documenti.

Siano  $t, w, d$  rispettivamente **topic**, **word**, **document**:

- Assegna casualmente ogni parola in ogni documento ad un topic
- Per ogni documento
  - Per ogni topic
    - Calcola la percentuale delle parole nel documento che sono assegnate al topic  $p(t | d)$
    - Per ogni parola nel documento
      - Calcola la percentuale delle assegnazioni al topic per tutti i documenti che contengono la parola:  $p(w | t)$
      - Moltiplica  $p(t | d) * p(w | t)$  ed assegna la parola ad un nuovo topic in base a questa probabilità
- Ripeti per un numero prefissato di iterazioni.





# Estrazione dei topic

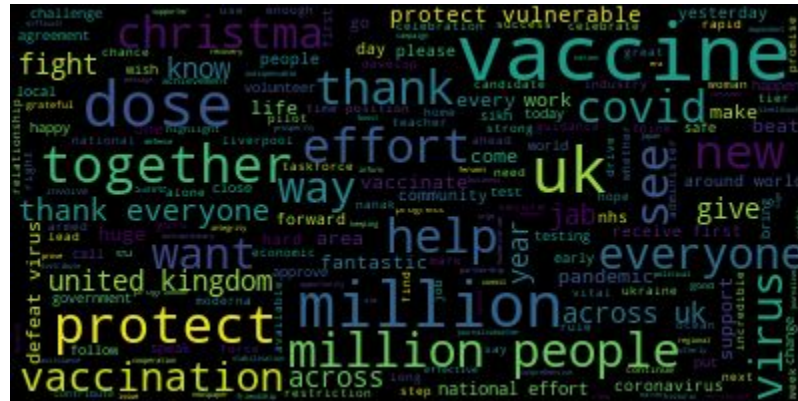
Per il rilevamento dei topic è stato effettuato:

- Vettorizzazione dei tweet
- Applicazione di LDA

Entrambe le operazioni sono state eseguite per mezzo della libreria Scikit-Learn.

# Creazione wordcloud

Raggruppati i tweet per topic, sono state generate le relative wordcloud, per evidenziare le parole più frequenti per ciascun topic.





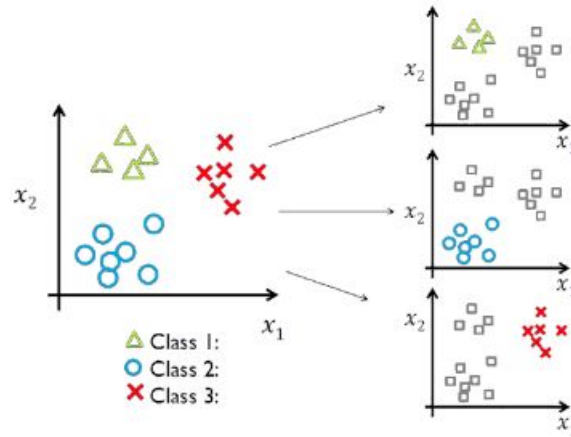
# Creazione modello di emotion detection

È stato creato un classificatore per effettuare l'emotion analysis dei tweet, adoperando come training set un dataset contenente circa 35000 tweet etichettati con le classi *joy*, *sadness*, *fear*, *anger*, *surprise*, *neutral*, *disgust*, *shame*.

Sul dataset di addestramento sono state effettuate le medesime operazioni di cleaning e preprocessing fatte sui dataset dei tweet estratti in precedenza, ed applicato il count vectorize.

Per la classificazione è stato utilizzato il Logistic Regression

# Logistic Regression



Accuracy media: 0.63



# Emotion analysis dei tweet

Il modello così salvato è stato utilizzato per stabilire l'emotion dei tweet non etichettati.

Per ciascun leader, sono state rilevate:

- Per ciascun periodo, le emotion di ciascun topic
- Le emotion di ciascun periodo



# Risultati

Per ogni Paese e per ciascuno dei quattro periodi presi in considerazione, vengono mostrati:

- Wordcloud dei topic
- Diagrammi a torta inerenti le emotion analysis dei tweet di ciascun topic per ciascun periodo
- Diagrammi a torta inerenti la emotion analysis dei tweet totali (indipendentemente dal topic) di ciascun periodo

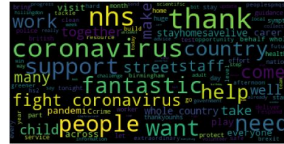
## List of topics 01/01/2020 - 31/05/2020 - United-Kingdom



Topic 1



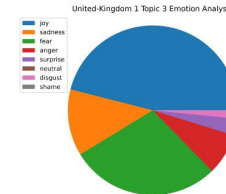
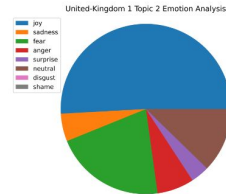
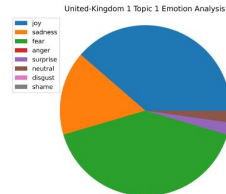
Topic 2



Topic 3



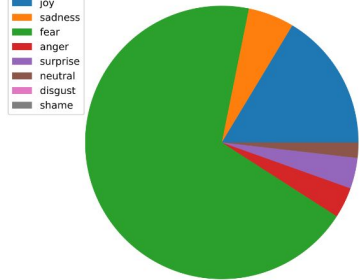
Emotion Pie Topic 1 Emotion Pie Topic 2 Emotion Pie Topic 3




A word cloud visualization of tweets from March 2020 related to COVID-19. The words are arranged in a circular pattern, with larger fonts indicating higher frequency. Key terms include "stay at home", "coronavirus", "beat virus", "continue", "work together", "follow the rules", "keep your distance", "must need", "please stay home", "government", "party", "business", "disease", "stand", "peak", "lead", "level", "country", "step", "build", "achieve", "new", "rule", "role play", "defeat", "tackle", "double", "let's say", "thank you", "leader", "agree", "fight", "follow", "keep", "go", "staying alert", "everyone", "safe", "past", "begin", "next way his", "together".

United-Kingdom 1 Topic 4 Emotion Analysis

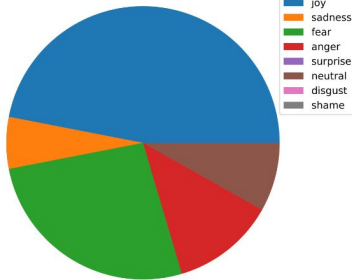
Emotion	Percentage
joy	45%
sadness	25%
fear	20%
anger	10%



United-Kingdom 1 Topic 5 Emotion Analysis

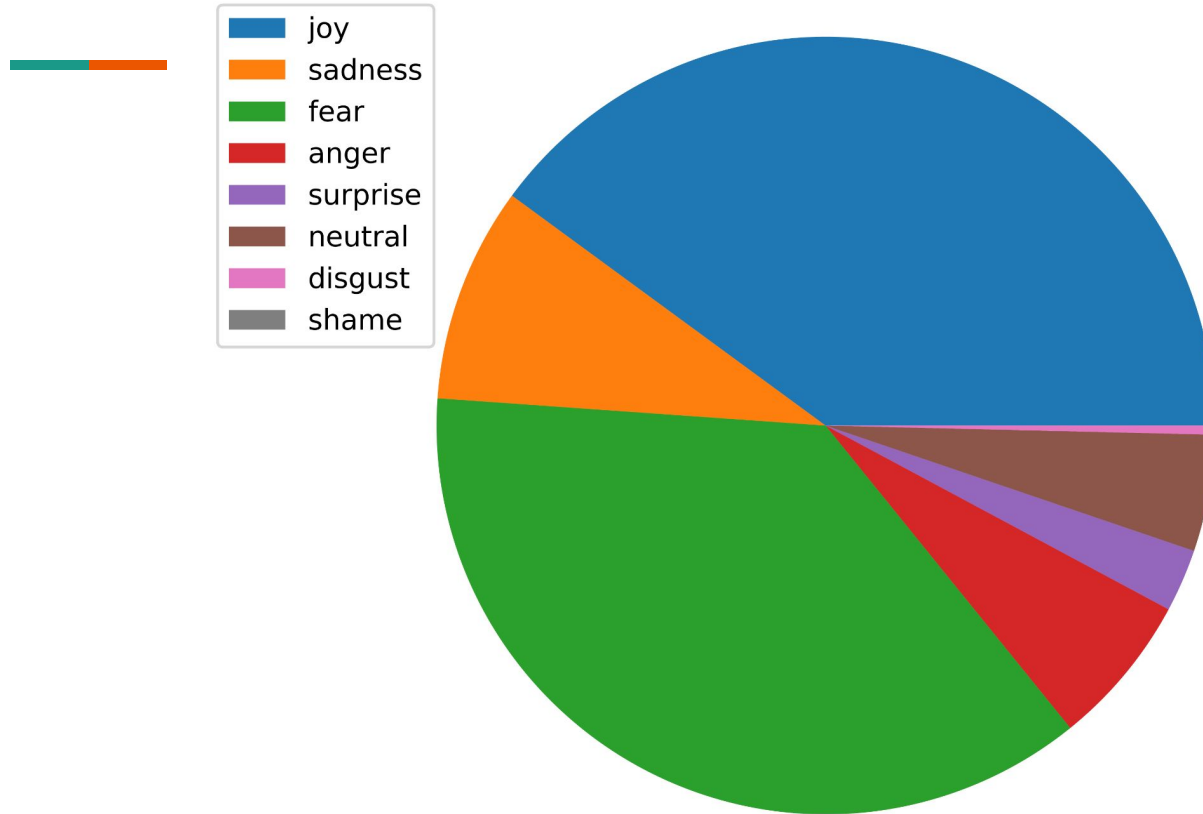


Emotion	Count
joy	10
sadness	0
fear	0
anger	0
surprise	0
neutral	0
disgust	0
shame	0





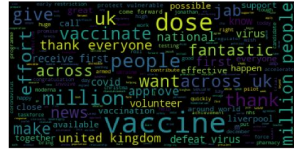
## United-Kingdom 01/01/2020 - 31/05/2020 Emotion Analysis



## List of topics 01/10/2020 - 31/12/2020 - United-Kingdom



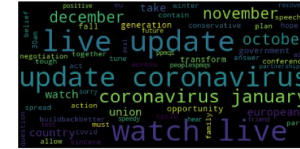
Topic 1



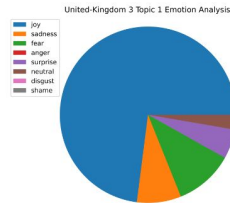
Topic 2



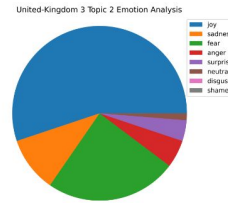
Topic 3



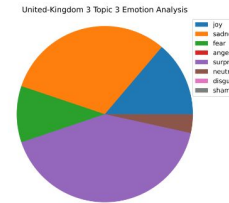
Emotion Pie Topic 1



Emotion Pie Topic 2

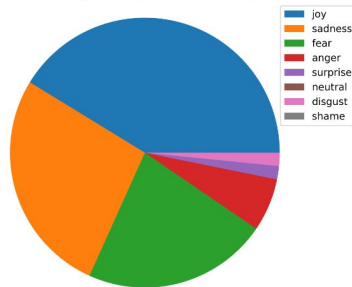


Emotion Pie Topic 3

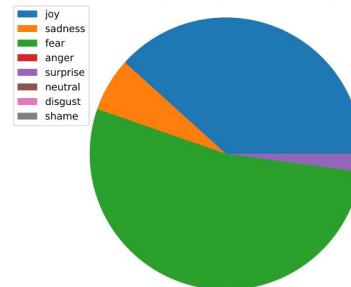


[illegible][illegible]

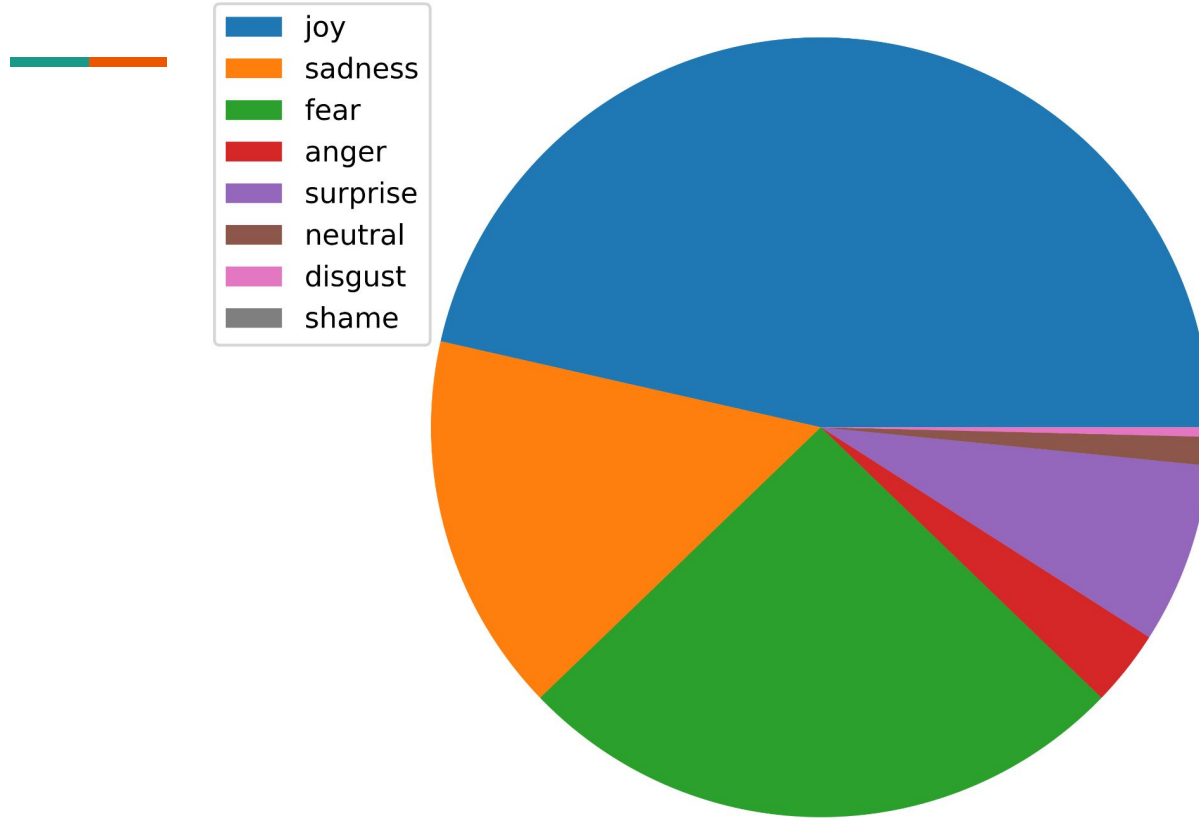
United-Kingdom 3 Topic 4 Emotion Analysis



United-Kingdom 3 Topic 5 Emotion Analysis



## United-Kingdom 01/10/2020 - 31/12/2020 Emotion Analysis



[illegible][illegible][illegible]

Brasil 2 Topic 1 Emotion Analysis

Emotion	Percentage
joy	35%
sadness	25%
fear	10%
anger	5%
surprise	2%
neutral	1%
disgust	1%
shame	1%

Brasil 2 Topic 2 Emotion Analysis

Emotion	Percentage
joy	45%
sadness	15%
fear	15%
anger	10%
surprise	5%
neutral	5%
disgust	2%
shame	2%

## List of topics 01/06/2020 - 30/09/2020 - Brasil



### Topic 4

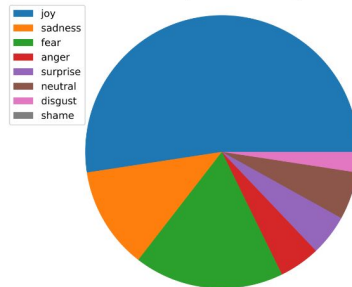


### Topic 5



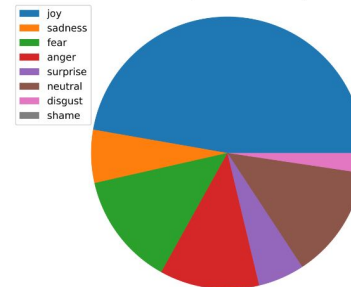
### Emotion Pie Topic 4

Brasil 2 Topic 4 Emotion Analysis

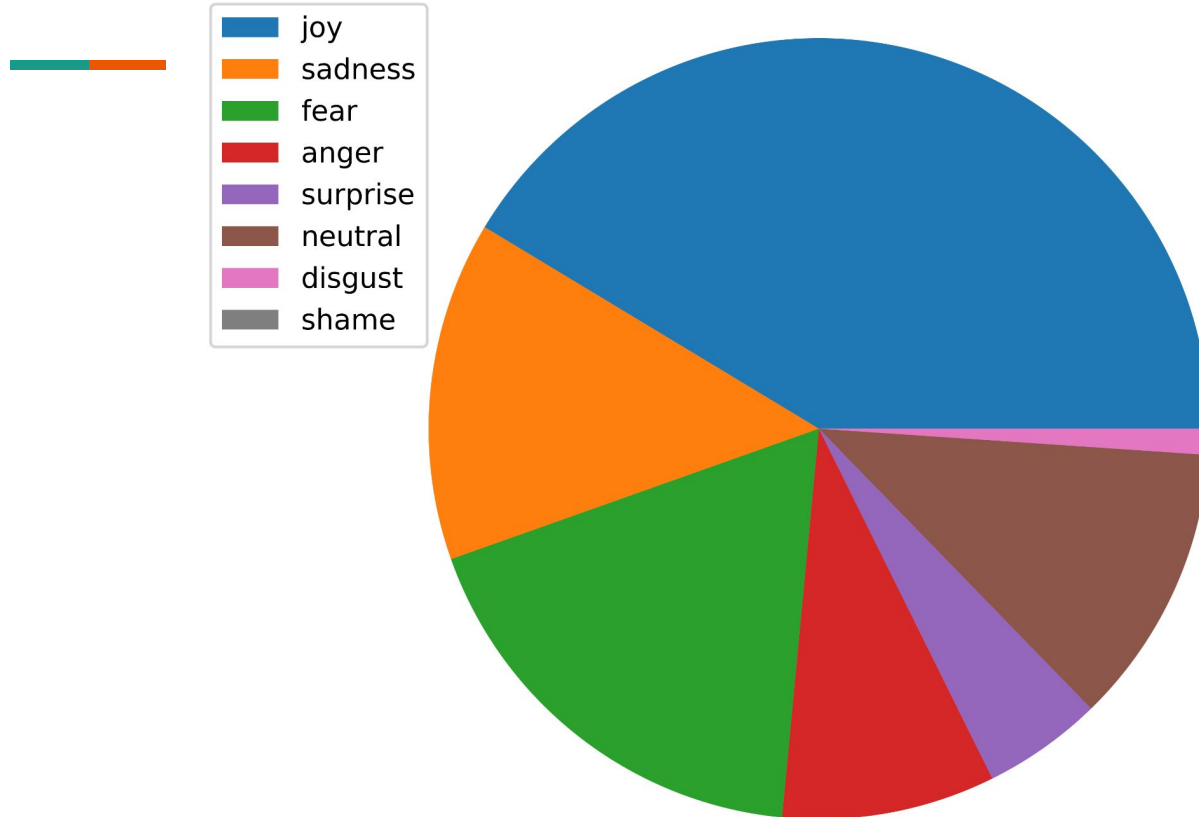


### Emotion Pie Topic 5

Brasil 2 Topic 5 Emotion Analysis



## Brasil 01/06/2020 - 30/09/2020 Emotion Analysis



[illegible]

cultura uva attivila torza  
misura decretare presidente  
**governare** milione tanto  
quattroprola zingari proutesse  
firmareprimogogliomontano  
azione lavorare partitire  
coronavirus pool dopo  
matina lipoartar e  
paese nuovo intervento  
bosniaca scibile

**Foto: Gettyimages.com - Contrasto / Ansa / Imagoeconomica**

[illegible]

Italy 1 Topic 2 Emotion Analysis

Emotion	Percentage
joy	45%
sadness	5%
fear	25%
anger	15%
surprise	5%
neutral	10%
disgust	0%
shame	0%

Italy 1 Topic 3 Emotion Analysis

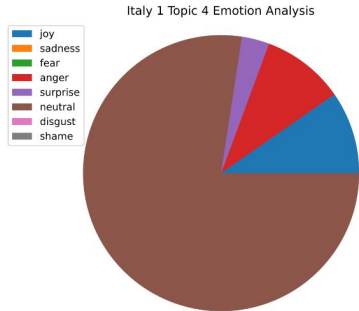
Emotion	Percentage
joy	45%
sadness	15%
fear	25%
anger	10%
surprise	5%
neutral	10%
dislike	10%
shame	5%



[illegible]

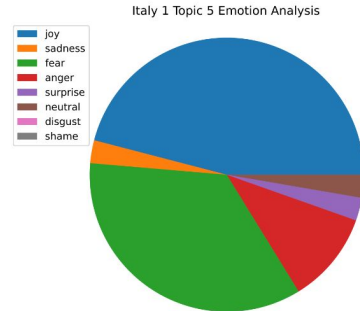
Italy 1 Topic 4 Emotion Analysis

Emotion	Percentage
joy	15%
sadness	10%
fear	5%
anger	10%
surprise	5%
neutral	45%
disgust	5%
shame	5%

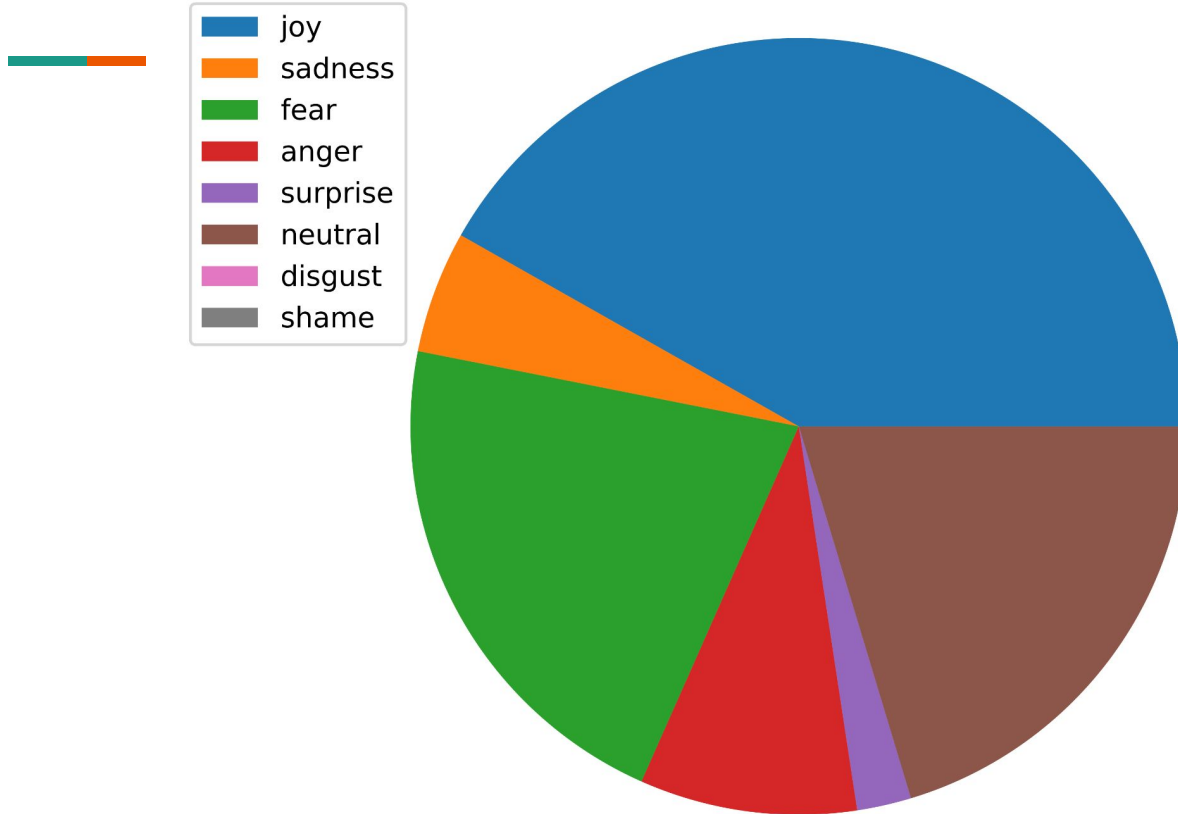


Italy 1 Topic 5 Emotion Analysis

Emotion	Percentage (approx.)
joy	45%
sadness	5%
fear	40%
anger	10%
surprise	2%
neutral	2%
disgust	1%
shame	1%



## Italy 01/01/2020 - 31/05/2020 Emotion Analysis

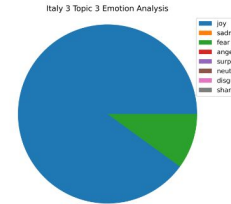
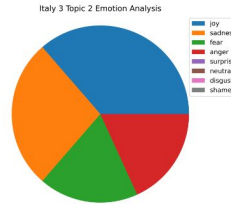
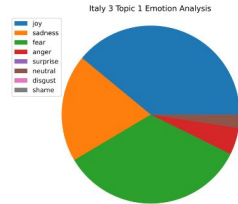


## Topic 2

[illegible]

## Emotion Pie Topic 2

### Emotion Pie Topic 3

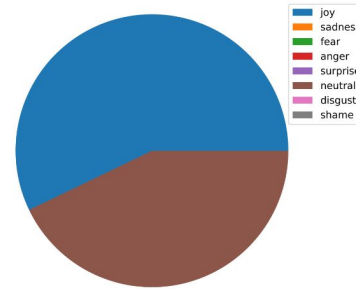




comunicazione appena concludere consueto  
 ministro senato mattina ordine  
 fino camera intervenire  
 associazione consigliare  
 collegamento poco madama  
 stampare chigi  
 situazione statigenerali55 villa luogo  
 palazzo conferenza  
 fra nazionale organizzato

## Emotion Pie Topic 5

Italy 3 Topic 5 Emotion Analysis



## Italy 01/10/2020 - 31/12/2020 Emotion Analysis

