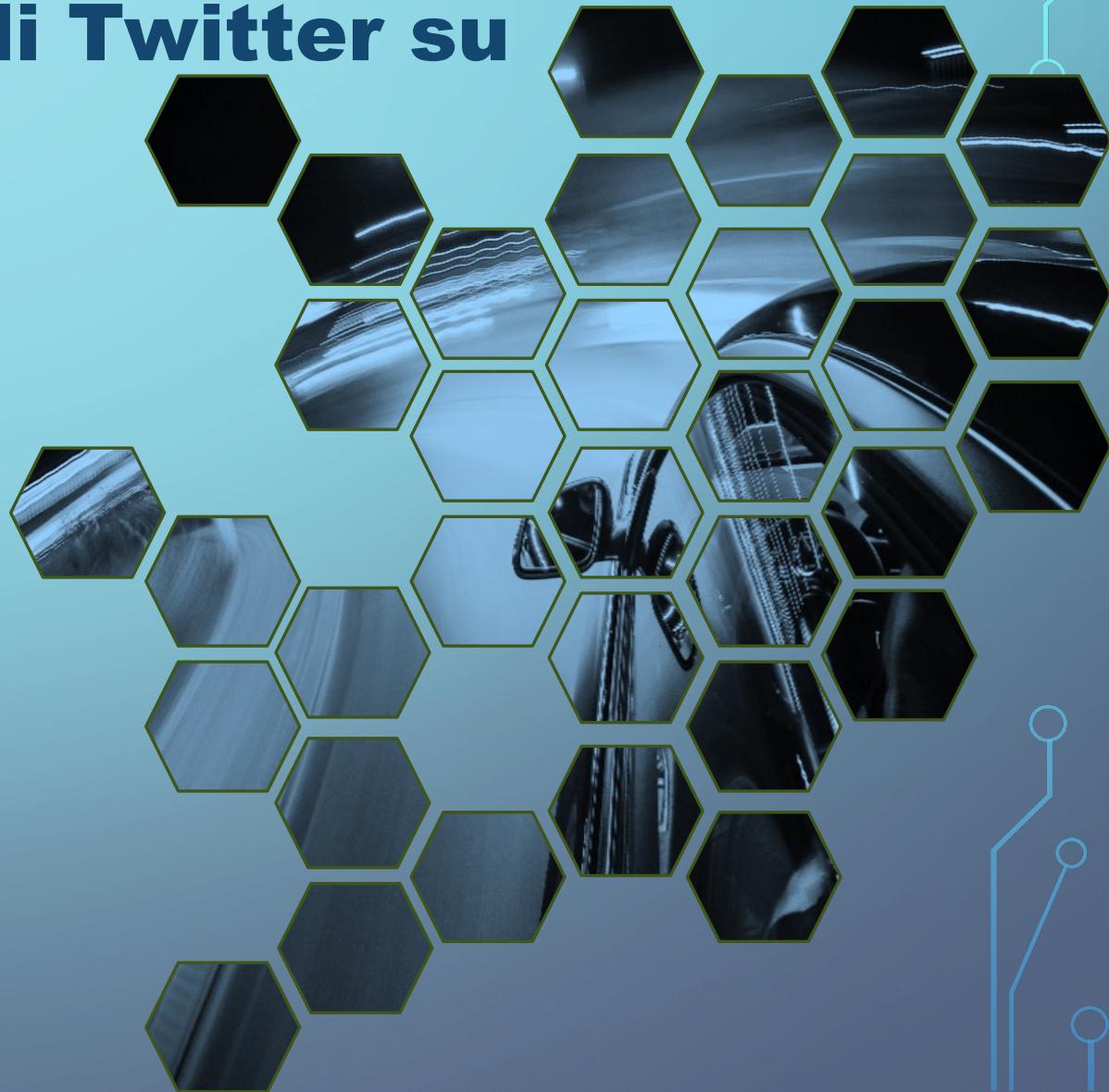


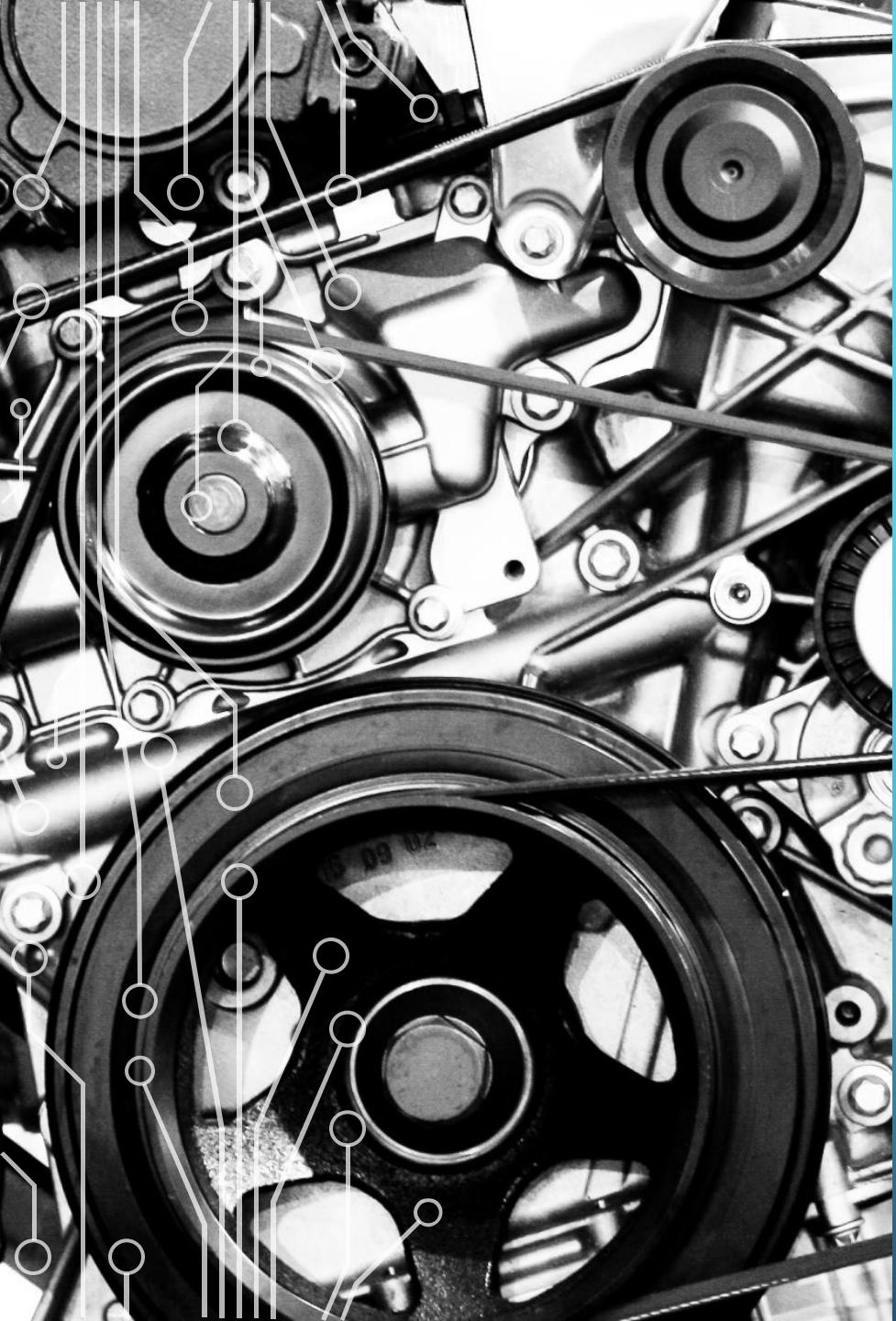
# **DriveFeelings: Cosa pensano gli utenti di Twitter su BMW, Renault e Tesla?**





# INTRODUZIONE

- Il progetto analizza i testi dei tweets degli utenti del social network Twitter che contengono le parole chiave di tre case automobilistiche scelte dai componenti del team e tra le più rilevanti del settore:
  - BMW
  - Tesla
  - Renault



## OBIETTIVI & MOTIVAZIONI

- La nostra passione per i motori ci ha spinto ad indagare:
- Cosa pensano gli utenti riguardo queste case automobilistiche, qual è la più apprezzata e la meno apprezzata.
- Capire gli argomenti a cui fanno riferimento gli utenti e i temi di discussione per ogni casa.
- Classificare i risultati per confrontare le differenze tra i soggetti dell'analisi.



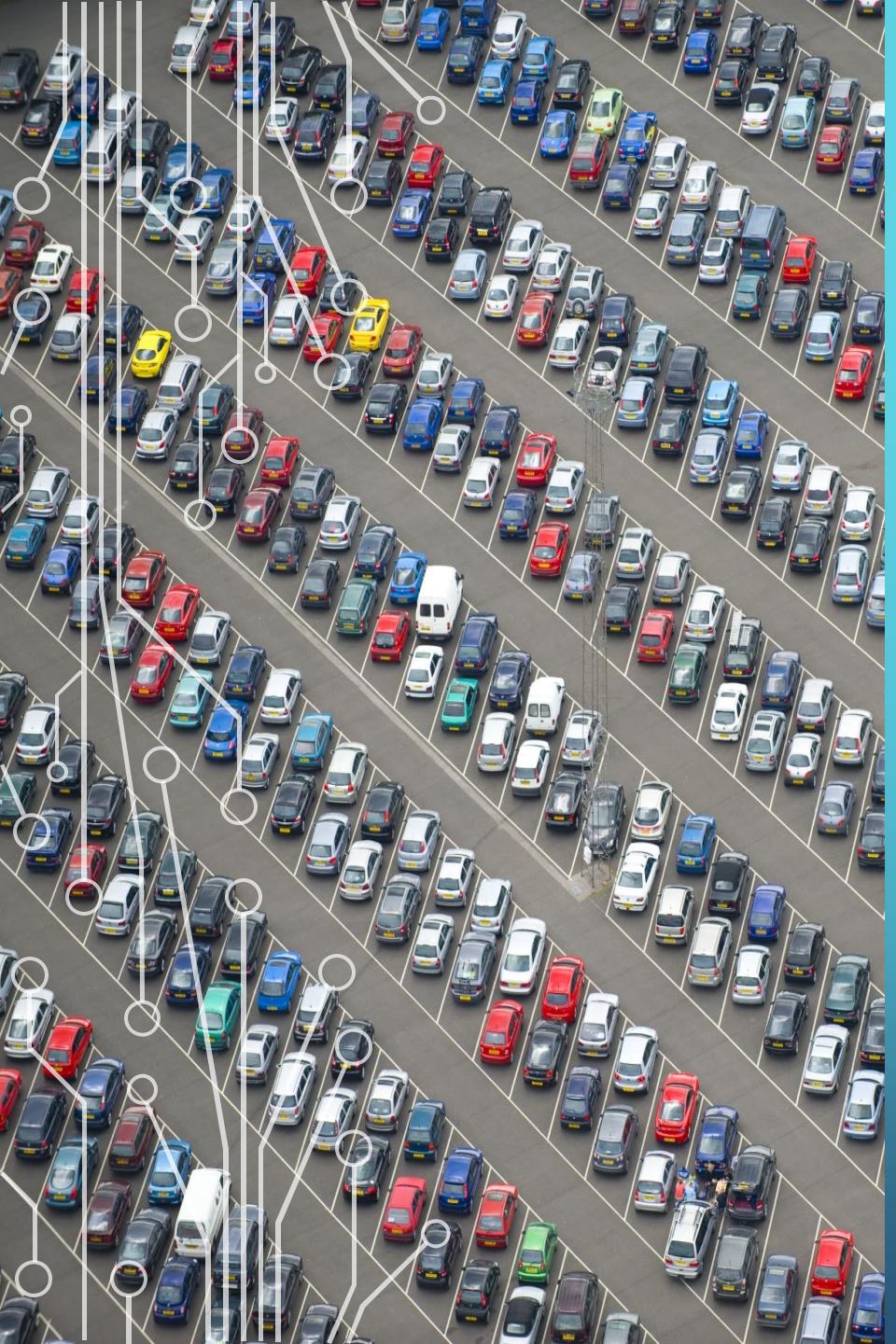
## FASI DELL'ANALISI:

- Twitter Scrapper
- Pre-Processing
- Topic Modeling
- RoBERTa SentimentAnalysis



## BRANCH 1: TWITTER SCRAPPER

- Estrazione dei tweets da Twitter attraverso uno scraper eseguito con twscrape
- Estrazione di 10.000 tweets per ogni casa automobilistica: BMW, Tesla e Renault
- Creazione di tre file distinti
- Totale di tweets estratti: 30.000



## BRANCH 2: PRE-PROCESSING

- I dati estratti contenevano rumore, informazioni inutili e testo sporco
- Sono state eseguite le seguenti operazioni:
  - Rimozione di caratteri speciali, punteggiatura, link e menzioni
  - Tokenizzazione del testo
  - Rimozione delle stopwords
  - Rimozione delle parole aggiuntive
  - Lemmatizzazione
- I dati sono stati pre-processati per affinare l'analisi



## BRANCH 3: TOPIC MODELING

- Per raggruppare i tweets in un numero ottimale di topic abbiamo usato la GridSearchCV
- Per la topic modeling abbiamo usato la LDA
- I topic identificati sono stati esplorati e analizzati per ottenere una visione dettagliata delle tematiche più frequenti tra gli utenti di Twitter
- Abbiamo creato dei grafici per la visualizzazione e l'interpretazione dei risultati



## BRANCH 4: ROBERTA SENTIMENTANALYSIS

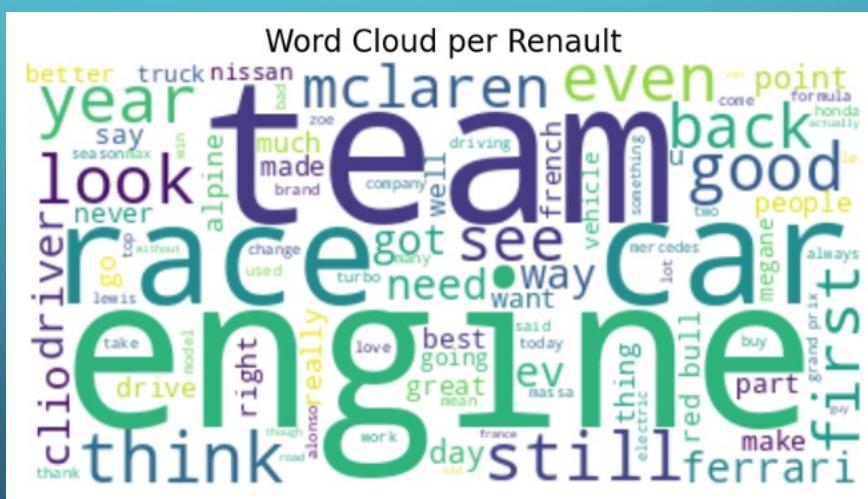
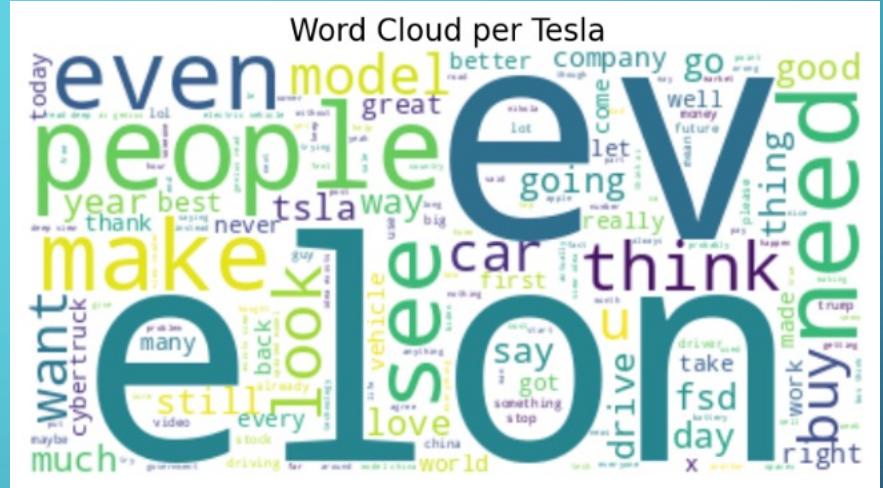
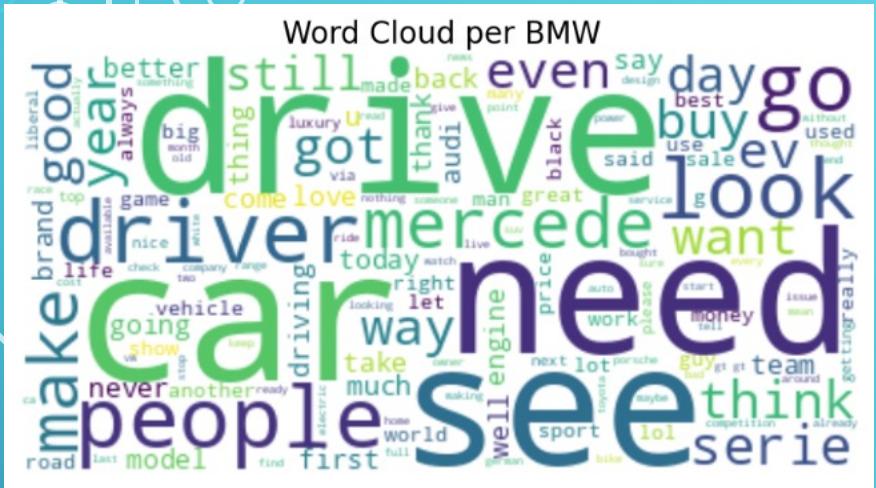
- L'analisi dei sentiment è stata eseguita sui dati pre-elaborati per determinare il sentiment generale dei tweet nei confronti delle tre case automobilistiche
- E' stato utilizzato un modello di apprendimento automatico (RoBERTa) per classificare i tweet in tre categorie principali: positive, neutral e negative confrontando l'accuratezza tra i modelli pre-trained e fine-tuned, così da etichettare i dati utilizzando il modello più performante.
- Questa analisi ha permesso di comprendere la percezione pubblica delle case automobilistiche sulla piattaforma Twitter



## INTERPRETAZIONE DEI RISULTATI

- Il progetto ha fornito una visione approfondita delle opinioni e delle discussioni online relative alle case automobilistiche
- Sono stati identificati i sentimenti prevalenti e i principali argomenti di discussione. L'analisi dei dati ha reso possibile una migliore comprensione della percezione degli utenti e ha fornito informazioni utili per identificare quali fossero i punti di forza e debolezza dei tre brand automobilistici

# GRAFICI WORD CLOUD



## 5 TOPIC RESULTS:

- RENAULT

Miglior Punteggio di verosimiglianza logaritmica (Log Likelihoods): -197360.97782693803  
Parametri del modello migliore: {'learning\_decay': 0.7, 'n\_components': 5}  
Perplexity del modello: 3697.502678420452  
Numero di Topics Ottimale: 5

Topic 1: engine, team, mclaren, ferrari, red, bull, good, mercedes, got, better  
Topic 2: win, race, year, think, want, point, truck, lewis, man, got  
Topic 3: ev, scenic, electric, car, year, look, range, nissan, battery, vehicle  
Topic 4: alonso, race, massa, grand, prix, make, people, driver, championship, hamilton  
Topic 5: alpine, team, gt, car, clio, drive, old, wheel, electric, van

- TESLA

Miglior Punteggio di verosimiglianza logaritmica (Log Likelihoods): -177669.40126407205  
Parametri del modello migliore: {'learning\_decay': 0.9, 'n\_components': 5}  
Perplexity del modello: 3357.2634947761535  
Numero di Topics Ottimale: 5

Topic 1: view, ai, think, idea, fsd, read, deep, genius, bos, good  
Topic 2: model, china, stock, ev, price, tsla, elon, cybertruck, cost, work  
Topic 3: electric, ev, vehicle, buy, battery, elon, need, way, good, truck  
Topic 4: love, company, got, make, delivery, car, year, tsla, ev, day  
Topic 5: people, need, want, elon, model, thing, ev, day, going, think

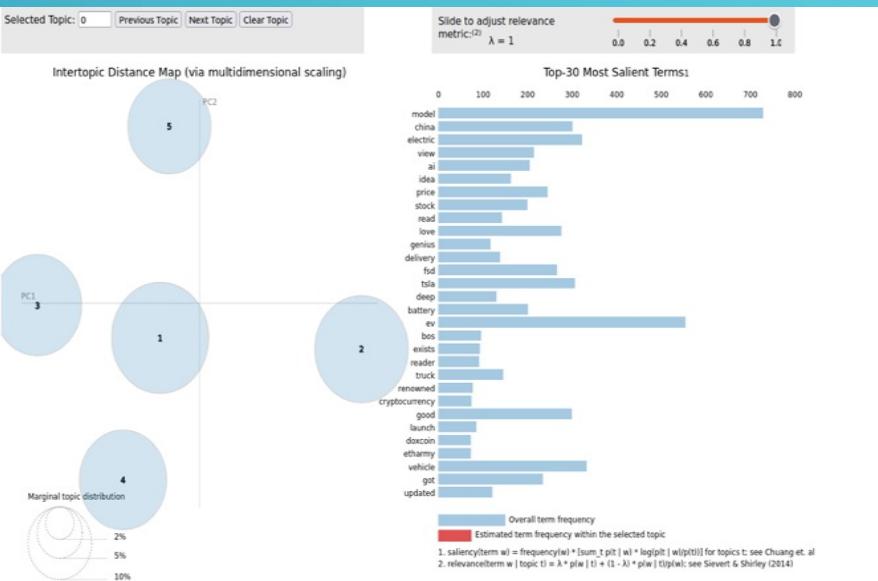
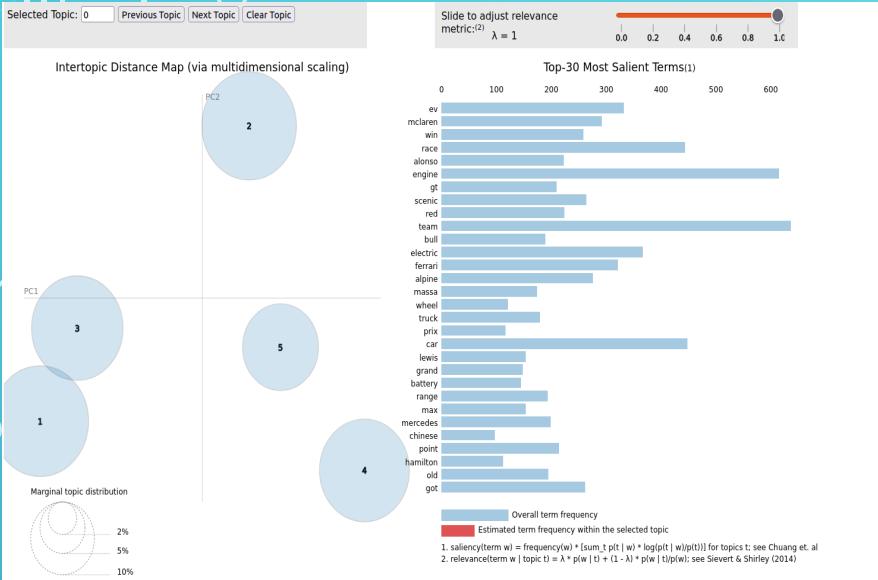
- BMW

Miglior Punteggio di verosimiglianza logaritmica (Log Likelihoods): -160005.30410910514  
Parametri del modello migliore: {'learning\_decay': 0.7, 'n\_components': 5}  
Perplexity del modello: 4340.666633156796  
Numero di Topics Ottimale: 5

Topic 1: gt, car, series, competition, year, company, sale, today, drive, day  
Topic 2: driver, got, guy, big, game, look, going, driving, fan, nice  
Topic 3: want, people, good, need, drive, audi, car, make, driving, love  
Topic 4: engine, mercedes, model, thank, benz, motorcycle, design, think, come, owner  
Topic 5: electric, ev, suv, india, year, black, range, luxury, vehicle, drive

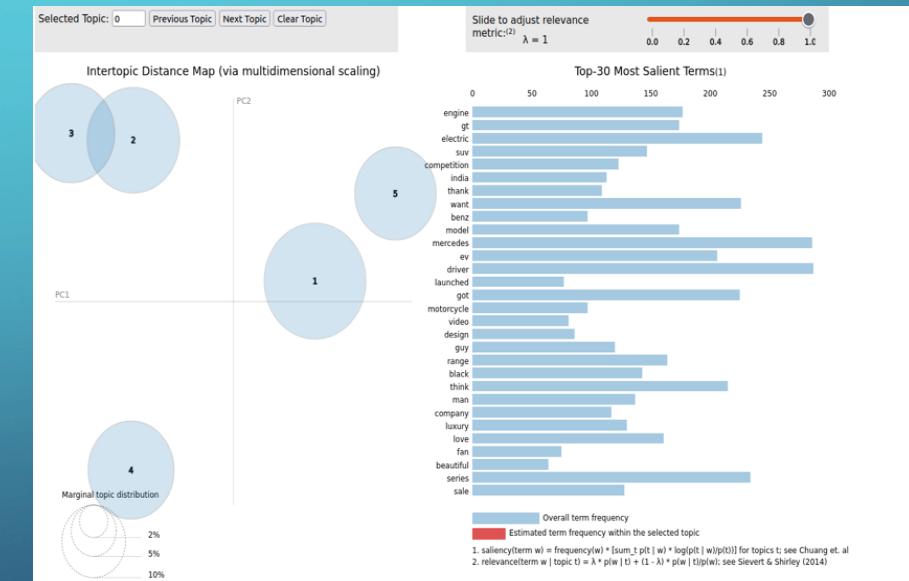
# GRAFICI PAGINE HTML:

1. RENAULT
2. TESLA
3. BMW



1

2

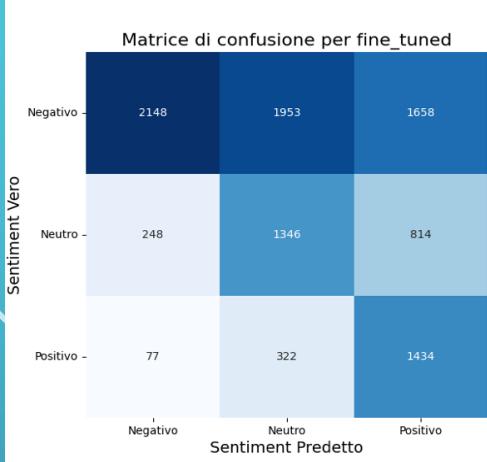


3

# ADDESTRAMENTO DEL MODELLO: FINE-TUNING

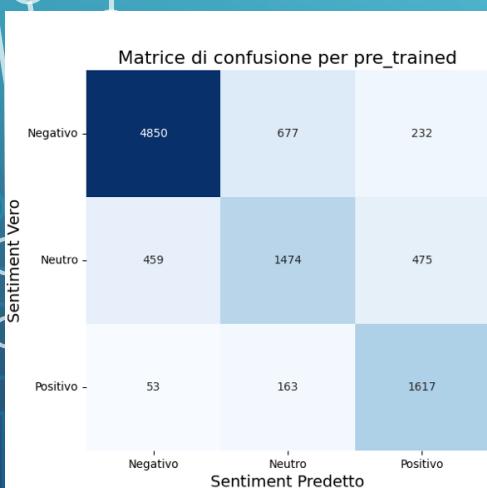
- Il fine tuning è stato eseguito sul modello RoBERTa pre-addestrato utilizzando il dataset Twitter\_data.csv
- Questo dataset presenta nella totalità 162980 tweets pre-etichettati, ma noi abbiamo optato di utilizzare solo i primi 30000 tweets perché ci siamo resi conto che computazionalmente è molto oneroso, di fatto ci ha messo circa 8 ore
- Per la validazione del modello abbiamo utilizzato i primi 10000 tweets del dataset: Test\_airlines\_sentiment.csv per calcolare l'accuracy del modello
- I dataset sono stati presi da Kaggle

# VALIDAZIONE E CONFRONTO DEI MODELLI



Accuracy su per modello fine\_tuned: 49.28%  
Confusion Matrix per modello fine\_tuned:  
[[2148 1953 1658]  
 [ 248 1346 814]  
 [ 77 322 1434]]  
RoBERTa Classification Report per fine\_tuned:

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| negative     | 0.87      | 0.37   | 0.52     | 5759    |
| neutral      | 0.37      | 0.56   | 0.45     | 2408    |
| positive     | 0.37      | 0.78   | 0.50     | 1833    |
| accuracy     |           |        | 0.49     | 10000   |
| macro avg    | 0.54      | 0.57   | 0.49     | 10000   |
| weighted avg | 0.66      | 0.49   | 0.50     | 10000   |



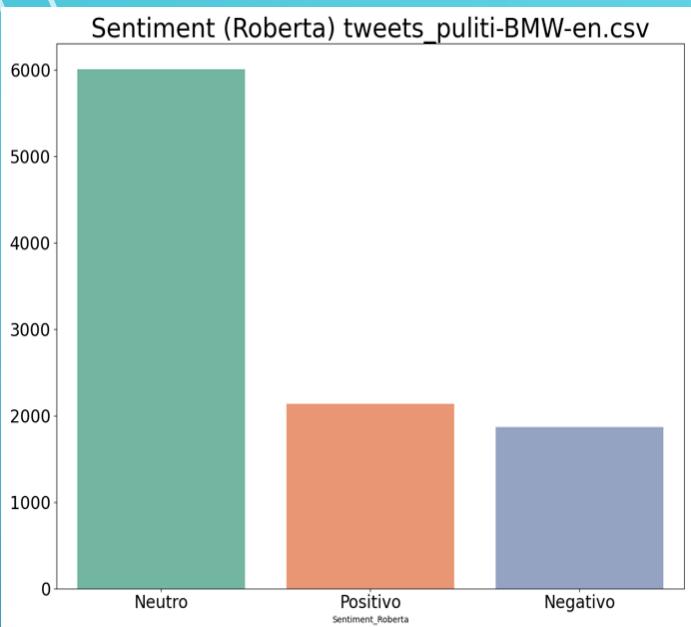
Accuracy su per modello pre\_trained: 79.41%  
Confusion Matrix per modello pre\_trained:  
[[4850 677 232]  
 [ 459 1474 475]  
 [ 53 163 1617]]  
RoBERTa Classification Report per pre\_trained:

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| negative     | 0.90      | 0.84   | 0.87     | 5759    |
| neutral      | 0.64      | 0.61   | 0.62     | 2408    |
| positive     | 0.70      | 0.88   | 0.78     | 1833    |
| accuracy     |           |        | 0.79     | 10000   |
| macro avg    | 0.75      | 0.78   | 0.76     | 10000   |
| weighted avg | 0.80      | 0.79   | 0.80     | 10000   |

**Modello fine-tuned:**  
Accuracy: 49,28%

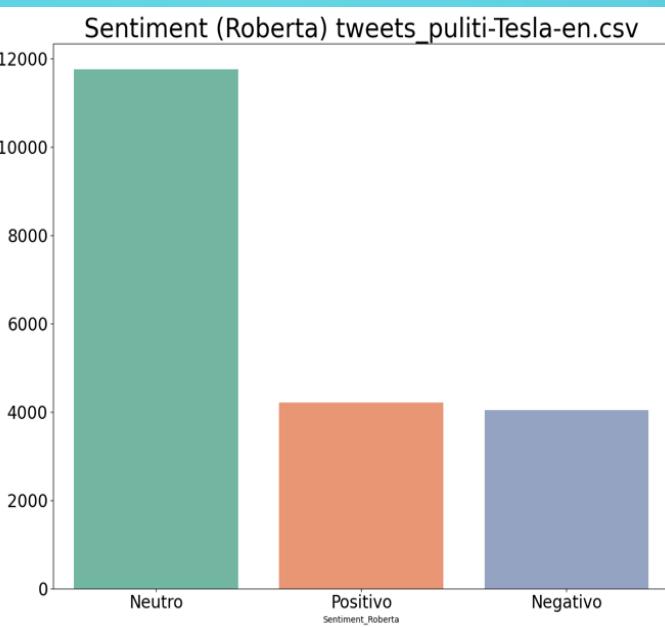
**Modello pre-trained:**  
Accuracy: 79,41%

# GRAFICI A BARRE: SENTIMENT ANALYSIS



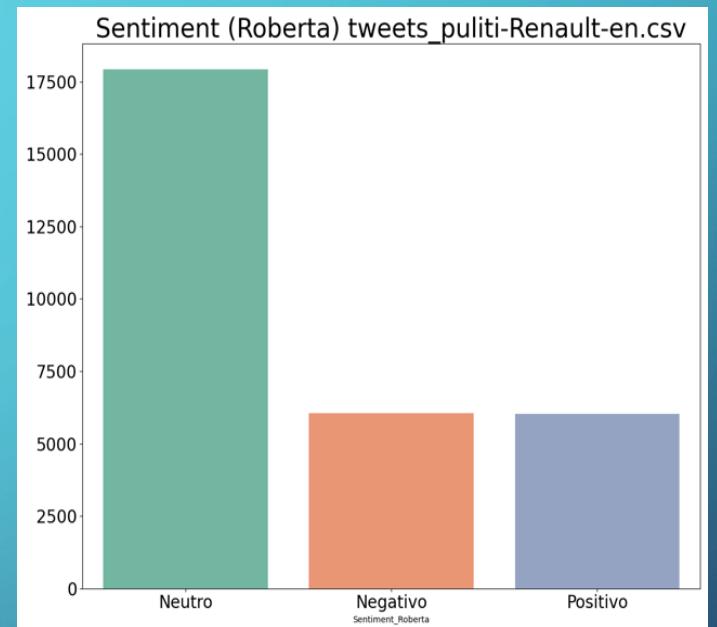
BMW

**Positivi: 21.34%**  
**Neutri: 60.02%**  
**Negativi: 18.64%**



TESLA

**Positivi: 21.04%**  
**Neutri: 58.78%**  
**Negativi: 20.18%**



RENAULT

**Positivi: 20.06%**  
**Neutri: 59.74%**  
**Negativi: 20.2%**

# CONCLUSIONI E PROBLEMI RISCONTRATI

- Il progetto ha fornito una visione approfondita delle opinioni e delle discussioni online relative alle case automobilistiche BMW, Renault e Tesla.
- Una considerazione sulla causa della presenza della dominanza del sentimento neutro è il fatto che molti tweets siano di natura pubblicitaria o prodotti da bot.
- La presenza dei bot è confermata dalle politiche che Elon Musk sta inserendo per il social network Twitter, cercando di limitare la presenza di profili non verificati e che producono spam inserendo delle licenze a pagamento per disincentivare la creazione di profili di questo tipo.
- Tuttavia, un aspetto importante che dobbiamo considerare è la presenza di classi sbilanciate all'interno dei dati. Questo significa che potrebbero esserci discrepanze significative nel numero di dati rappresentativi per ciascuna casa automobilistica. Affrontare questo problema è cruciale per garantire risultati più accurati e bilanciati. Una strategia per far fronte alle classi sbilanciate potrebbe essere l'utilizzo di tecniche di bilanciamento dei dati, come l'oversampling o l'undersampling. Queste metodologie consentono di equilibrare il dataset, garantendo che ciascuna casa automobilistica abbia un peso paritario nell'analisi.