Seminario: A Caccia di Fake News

PCTO - Summer Camp Informatica & Intelligenza Artificiale

Giovedì, 4 Settembre 2025

Michael Soprano - michael.soprano@uniud.it

Università degli Studi di Udine

Scaletta

1. Che cos’è (e non è) una fake news
2. Fact-checking e ground truth
3. Accordo: misurare il consenso
4. Mini-demo: fact-checker automatico

Chi sono

Michael Soprano

Ricercatore @ Università degli Studi di Udine

Dipartimento di Scienze Matematiche, Informatiche e Fisiche (DMIF)

Lab. Social, Mobile, Data & Crowd (SMDC)

Email: `michael.soprano@uniud.it`

Homepage: www.michaelsoprano.com

Cominciamo!

- Questa immagine è **autentica?**

Spoiler: è un falso

- **Immagine virale** a marzo 2023
- **Papa Francesco** con un piumino bianco
- **Ampia diffusione** tra celebrità, influencer e alcune testate online
- **Generata con intelligenza artificiale (IA)**

Cos’è una “fake news”?

> *Contenuto falso o fuorviante presentato come notizia di attualità*

- È un’**espressione ombrello**, ampia e **imprecisa**
- Servono **termini più specifici** e **definizioni rigorose**

Usiamo termini più precisi

- **Misinformazione**: falsa, diffusa **senza intento** di ingannare
  - Es.: storia Instagram con un dato sbagliato condiviso in buona fede
- **Disinformazione**: falsa, diffusa per **ingannare**
  - Es.: sito di notizie inventate creato per ottenere click o fare propaganda
- **Malinformazione**: vera, diffusa **per ingannare** o **fuori contesto**
  - Es.: pubblicazione di dati privati rilanciata come recente

---

Fonte: *Information Disorder* (Council of Europe, 2017)

Dove si incontrano online?

- **Facebook**: post sensazionalistici e titoli ingannevoli
- **WhatsApp**: audio “inoltrato molte volte” che rimbalzano tra gruppi
- **Instagram**: meme manipolati o ritagliati fuori contesto
- **TikTok**: video decontestualizzati o rimontati senza riferimento alla fonte

Social media in Italia

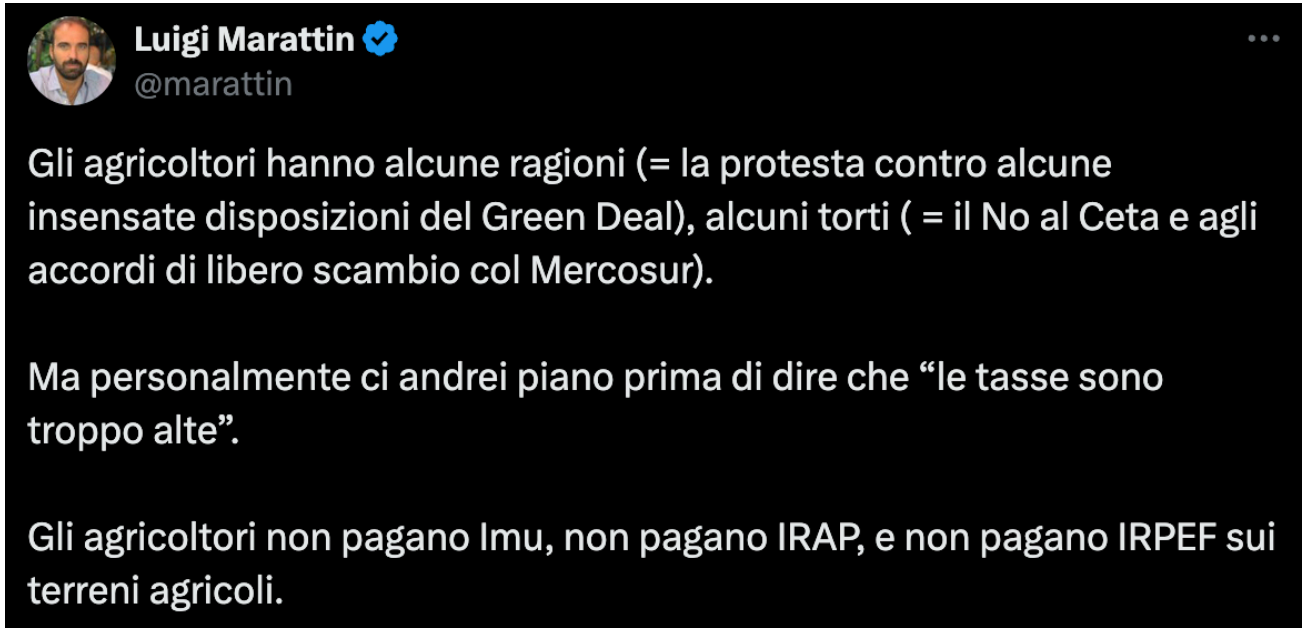
- ≈ **90%** della popolazione è online

Piattaforma	Utenti	Tempo/mese
TikTok	19,8 M (18+)	**≈ 30 h**
YouTube	42,2 M	17 h
Instagram	27,8 M	15 h

Fonte: *Digital 2025* (We Are Social & Meltwater / DataReportal) + ANSA, 27 feb 2025

Esempio reale — VERO

> *Gli agricoltori non pagano Imu, non pagano Irap, e non pagano Irpef sui terreni agricoli*



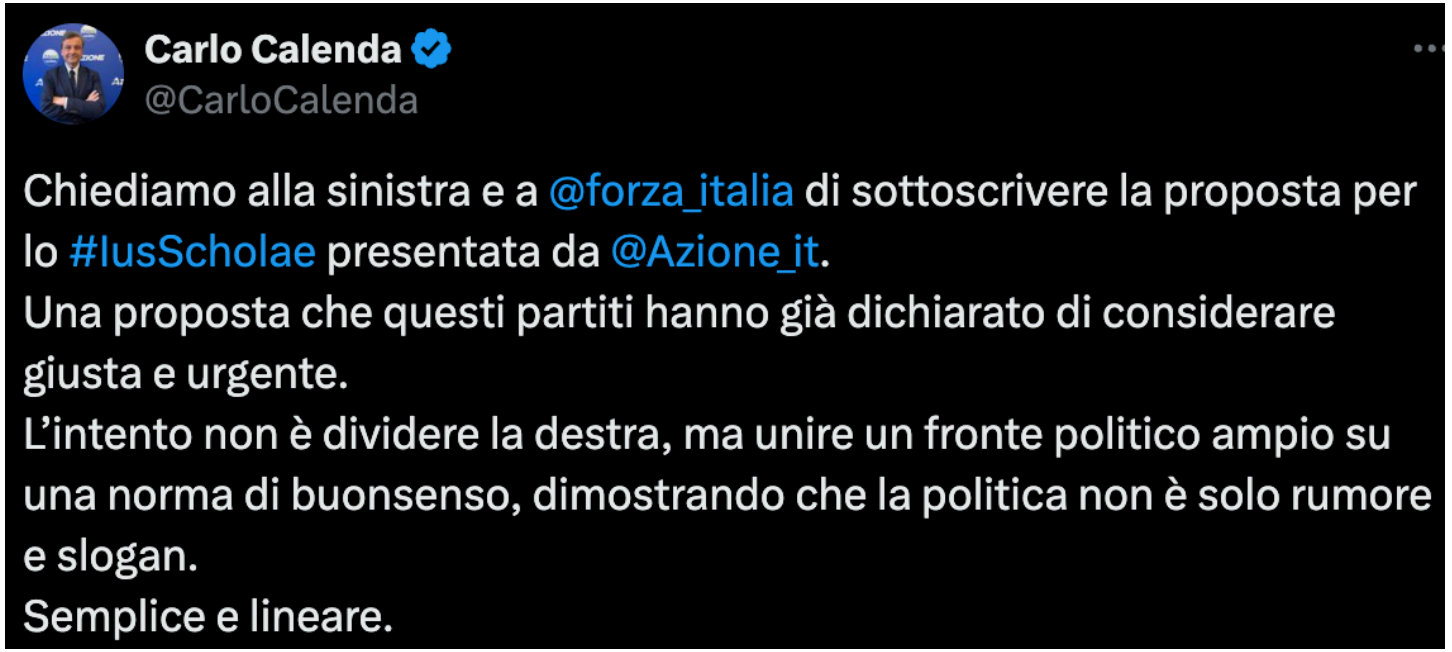
https://pagellapolitica.it/fact-checking/agricoltori-esenzione-imu-irap-irpef

Perché è corretto?

- **IMU**: esenti i terreni di coltivatori diretti e imprenditori agricoli professionali
- **IRAP**: esclusa l'attività agricola ai sensi di legge
- **IRPEF**: esenzione sui redditi agrari dal 2024 **non è più totale**
- **Pagella Politica**: verdetto VERO (6 febbraio 2024)

Esempio reale — FALSO

> *Siamo un Paese vecchio, dove in metà delle regioni i pensionati superano i lavoratori*



https://pagellapolitica.it/fact-checking/numero-pensionati-lavoratori-regioni

Perché è errato?

- **ISTAT 2022**: in **nessuna** regione italiana i pensionati superano i lavoratori
- Rapporto “peggiore”: **~1,8 lavoratori per ogni pensionato**
- **Pagella Politica** ha giudicato l’affermazione FALSA (29 agosto 2024)

Fonte dati e dettagli: Pagella Politica su dati ISTAT 2022

Cosa si può fare?

- Consultare organizzazioni di **fact-checking** come **Pagella Politica** e **PolitiFact**
- Queste realtà **verificano affermazioni pubbliche** e pubblicano **verdetti** basati su **fonti documentate**
- I loro archivi possono costituire una ***ground truth*** per **addestrare** e **valutare** gli algoritmi

How PolitiFact selects facts to check: https://www.youtube.com/watch?v=JDEFHkUNapo

Perché l’informatica aiuta

- I contenuti online sono troppi per una **verifica manuale**
- Algoritmi di IA possono **segnalare** contenuti sospetti e **prioritizzare** le verifiche
- Il giudizio finale resta **umano**: l’IA riduce il lavoro ripetitivo

Cos’è la «ground truth»?

- Insieme di esempi **verificati** e **etichettati** da fonti affidabili (es. fact-checker)
- È l’insieme di dati che **si sanno corretti** e si usa come **riferimento** in ricerca
- I fact-check pubblicati sono **strutturati**, **ricercabili** e **leggibili dalle macchine**
- Questi dati servono per **addestrare** modelli e per **valutare** le loro prestazioni

Come funziona in pratica

1. Si raccolgono **affermazioni pubbliche** da media, social o dichiarazioni ufficiali
2. I **fact-checker** analizzano l'affermazione e assegnano un'etichetta (es. `VERO` / `FALSO`)
3. Un **modello** di IA si addestra su questi **dati etichettati**
4. Il modello poi può **segnalare** nuove affermazioni **potenzialmente false**

Accordo tra valutatori

- Ogni affermazione è valutata da **almeno due** fact-checker
- Ciascun valutatore assegna **in modo indipendente** un'etichetta
- Se danno **sempre** la stessa risposta → **accordo = 100%**
- Se le **valutazioni divergono** spesso, conviene **analizzare le cause**
  - Frasi ambigue? Dati rumorosi? Istruzioni poco chiare?

Accordo percentuale (1/2)

- In inglese: *percent agreement*
- Misura quante volte i valutatori danno la **stessa risposta**

Affermazione	Valutatore A	Valutatore B
1	✅ Vero	✅ Vero
2	❌ Falso	❌ Falso
3	✅ Vero	❌ Falso
4	❌ Falso	❌ Falso
5	✅ Vero	✅ Vero

Accordo percentuale (2/2)

- Etichette uguali = **4 su 5 affermazioni**
- L'**accordo** tra i due valutatori è:

$$\text{Percent agreement} = \frac{4}{5} = 80\%$$

Come interpretarlo

- **100%** → i due valutatori danno **sempre** la stessa risposta
- **0%** → non sono **mai** d’accordo
- Valori intermedi indicano la **frequenza media di accordo**

Limiti e alternativa

- Se il **90%** delle frasi è `FALSO`, rispondere sempre `FALSO` dà comunque **90% di accordo**
- Il *percent agreement* misura solo la **corrispondenza**, non la **qualità** dell'accordo
- Alternativa: **κ di Cohen (kappa)**, che **corregge per l'accordo atteso per caso**

Il compito

- Verificare se la **fonte conferma** l'**affermazione**
- Un modo per **automatizzare** parte del fact-checking
- Esito: **confermata** o **non confermata**
- Misura: **accordo percentuale** tra sistema e **gold** (giudizio umano)

Il processo

1. Raccogliere esempi con **affermazione**, **fonte** e **gold** $\in$ {`VERO`, `FALSO`}
2. Confrontare **affermazione** e **fonte** con il sistema
3. Ottenere l'esito: **confermata** o **non confermata**
4. Confrontare con il **gold** e misurare l'**accordo**
5. Se l'accordo è basso, migliorare **esempi** e **metodo**

Come provare

- Eseguibile su **Google Colab**
- **Dataset** già nel notebook, senza file esterni
- Si possono **modificare affermazione e fonte** e osservare l'**accordo** in tempo reale
- Notebook: https://bit.ly/45Q6r16

Il dataset

- Ogni riga ha **3 campi**: `claim` (affermazione), `evidence` (fonte), `gold` (giudizio umano)
- Due versioni: **BASE** (fonte chiara) e **HARD** (fonte ambigua)

claim	evidence	gold
Torino è il capoluogo del Piemonte	Il capoluogo del Piemonte è Torino	VERO
Milano è la capitale d'Italia	La capitale d'Italia è Roma	FALSO

Il modello

- `MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7`
- Etichette NLI → **mappatura binaria**:
  - `ENTAILMENT` → **VERO (confermata)**
  - `CONTRADICTION` **o** `NEUTRAL` → **FALSO (non confermata)**

Passo 1 – Importiamo ciò che serve

- Importare le librerie per l'**analisi** e per il **modello**

```
from transformers import pipeline                # modello NLI
from sklearn.metrics import accuracy_score       # percent agreement vs gold
```

Passo 2: Definiamo il dataset

- **Due insiemi:** `BASE` e `HARD`

```
data_base = [
    ("Torino è il capoluogo del Piemonte", "Il capoluogo del Piemonte è Torino", True),   # vera
    ("Roma è la capitale d'Italia", "La capitale d'Italia è Roma", True),                 # vera
    # ...
]

data_hard = [
    ("Torino è il capoluogo del Piemonte", "Il capoluogo del Piemonte è Novara", False),  # contraddice
    ("Roma è la capitale d'Italia", "Roma è la città più popolosa d'Italia", False),      # non confermata (NEUTRAL)
    # ...
]
```

Passo 3 – Carichiamo il modello

```
from transformers import pipeline

clf = pipeline(
    task="text-classification",
    model="MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7",
    tokenizer="MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7",
    truncation=True
)
# Etichette possibili: ENTAILMENT / CONTRADICTION / NEUTRAL
# NLI: evidence = "premessa" → text ; claim = "ipotesi" → text_pair
```

Passo 4: Facciamo lavorare l'IA

- Due **colonne di etichette**: `gold` (umano) e `pred` (IA)

```python
def percent_agreement(dataset):
    # dataset: lista di tuple (claim, evidence, gold_bool)
    inputs = [{"text": ev, "text_pair": cl} for (cl, ev, _) in dataset]
    # Chiamata al modello di IA
    out = clf(inputs)

    # Mapping binario: True se l'etichetta inizia con "ENTAIL" (copre ENTAIL / ENTAILMENT)
    y_pred = [str(p.get("label", "")).upper().startswith("ENTAIL") for p in out]
    # Gold umano già "binarizzato": True=confermata, False=non confermata
    y_true = [g for (_, _, g) in dataset]

    # Match punto-a-punto tra predizioni e gold
    hits = sum(a == b for a, b in zip(y_true, y_pred))
    n = len(dataset)  # si assume n > 0 nella demo
    # Ritorna: (percent agreement / accuracy, numero di match, numero di esempi)
    return hits / n, hits, n
```

Passo 5: Calcoliamo l’accordo

- Confronto **sistema vs gold** su `BASE` e `HARD`

```
acc_b, hits_b, n_b = percent_agreement(data_base)
acc_h, hits_h, n_h = percent_agreement(data_hard)

print(f"BASE → accordo sistema vs gold: {hits_b}/{n_b} = {acc_b:.0%}")
print(f"HARD → accordo sistema vs gold: {hits_h}/{n_h} = {acc_h:.0%}")
# Differenza di accordo
print(f"Δ accordo (BASE - HARD): {(acc_b - acc_h):.1%}")
```

Extra: provare combinazioni diverse

- Esempi di **fonti** diverse per la stessa **affermazione**
- Vedere quando il modello **cambia risultato**

```
# Caso 1: fonte che conferma (atteso: ENTAILMENT → VERO)
print(clf([{
    "text": (
        "Lo Statuto regionale e atti amministrativi indicano Firenze come capoluogo della Toscana; "
        "non risultano delibere che ne abbiano disposto la modifica."
    ),
    "text_pair": "Firenze è il capoluogo della Toscana."
}])[0])
# Caso 2: fonte che contraddice (atteso: CONTRADICTION → FALSO)
print(clf([{
    "text": "Documenti ufficiali riportano Pisa come capoluogo regionale.",
    "text_pair": "Firenze è il capoluogo della Toscana."
}])[0])
```

Combinazioni affermazione/fonte

- **Affermazione fissa:** “Firenze è il capoluogo della Toscana”

Caso	Fonte	NLI	Binario
1	Statuto e atti amministrativi indicano Firenze come capoluogo; nessuna delibera di modifica	ENTAILMENT	VERO
2	Documenti ufficiali riportano Pisa come capoluogo regionale	CONTRADICTION	FALSO
3	Sede della giunta a Firenze; funzioni diffuse sul territorio	NEUTRAL	FALSO
4	Firenze fu capitale del Regno d’Italia nel XIX secolo	NEUTRAL	FALSO

Cosa portiamo a casa

1. Anche un'**immagine virale** può ingannare → serve **spirito critico**
2. Il lavoro dei fact-checker definisce la ***ground truth*** di riferimento
3. L'**accordo percentuale** misura quanto spesso **valutatori** coincidono
4. Con un **modello preaddestrato** si costruisce un **semplice supporto** al fact-checking
5. Un sistema ben progettato **affianca** il lavoro umano, non lo sostituisce

Risorse e approfondimenti

- Fact-checking in Italia: pagellapolitica.it
- Fact-checking negli USA: politifact.com
- Dataset internazionali: Google Fact Check Explorer