Giacomo Falcone

PROGETTO DI INGEGNERIA INFORMATICA

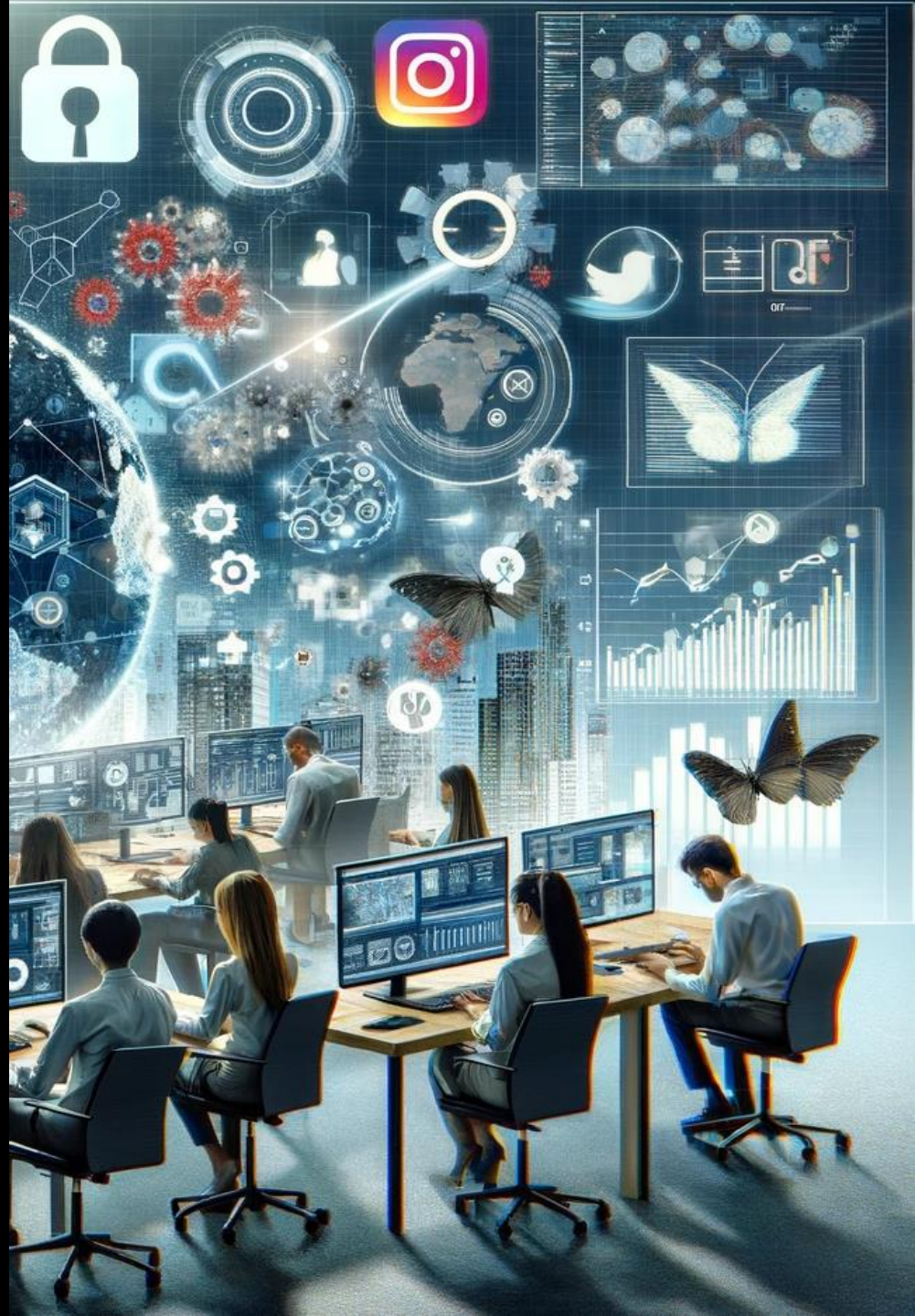
Utilizzo dei dati dei social media per l'analisi del sentiment e la previsione dei crolli delle criptovalute

A.A. 2023/2024

INTRODUZIONE E CONTESTO

Le criptovalute, come Bitcoin, sono diventate strumenti di investimento e speculazione riconosciuti a livello globale, caratterizzati da estrema volatilità. Questa volatilità rende il mercato altamente imprevedibile e influenzato da fattori come notizie di regolamentazioni e sentiment degli investitori.

Il progetto mira ad analizzare i dati dei social media per identificare possibili segnali di allarme che potrebbero preannunciare un futuro crollo di una criptovaluta.



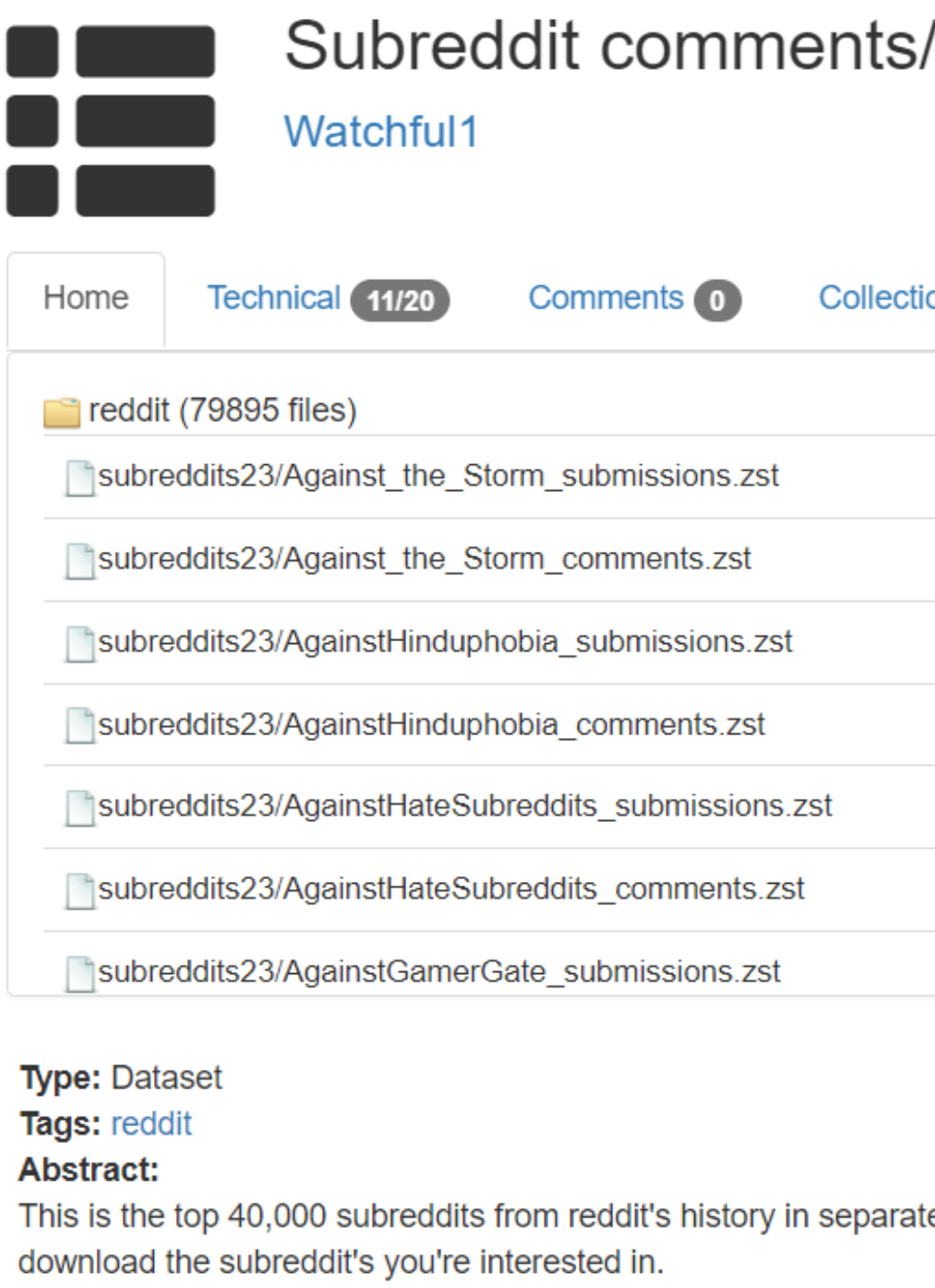
SCOPO DEL LAVORO

Negli ultimi anni, l'interesse per la previsione dei movimenti di mercato delle criptovalute è cresciuto. Sono disponibili molte sorgenti di informazione per descrivere e analizzare il mercato (notizie finanziarie, dati storici…). Tra queste i social media rappresentano una buona fonte di dati poiché riflettono il sentiment e le reazioni immediate degli investitori.

Come casi di studio, sono stati utilizzati i crolli di FTX e Terraform Labs. L'analisi si è concentrata quindi sui dati di post e commenti del subreddit CryptoCurrency (integrati con Google Trends.) Infine è stato sviluppato un modello di machine learning per prevedere i prezzi delle criptovalute, ponendo le basi per uno strumento prezioso per gli investitori.

METODO

1. Raccolta dei dati da fonti rilevanti (social media, Google Trends)
2. Pre-elaborazione dei dati per garantire qualità e coerenza (inclusi la pulizia, il filtraggio e la normalizzazione)
3. Analisi esplorativa dei dati per comprendere le tendenze e identificare eventuali pattern
4. Utilizzo di visualizzazioni per interpretare meglio i dati raccolti
5. Suddivisione dei dati in set di training e test per addestrare modelli di machine learning (regressione Random Forest)
6. Utilizzo di metriche di valutazione (come MSE e R^2) per misurare la precisione e la robustezza delle previsioni
7. Implementazione del modello per fare previsioni sui dati nuovi e futuri



DATI

Raccolta dei dati:

- Estrazione dei dati (raccolti da Pushshift) tramite Academic Torrents e filtraggio (intervallo di tempo) con script Python
- Utilizzo dei dati del subreddit CryptoCurrency (circa 100.000 commenti e 5.000 post (periodo di un mese intorno all'evento d'nteresse) per singolo caso)

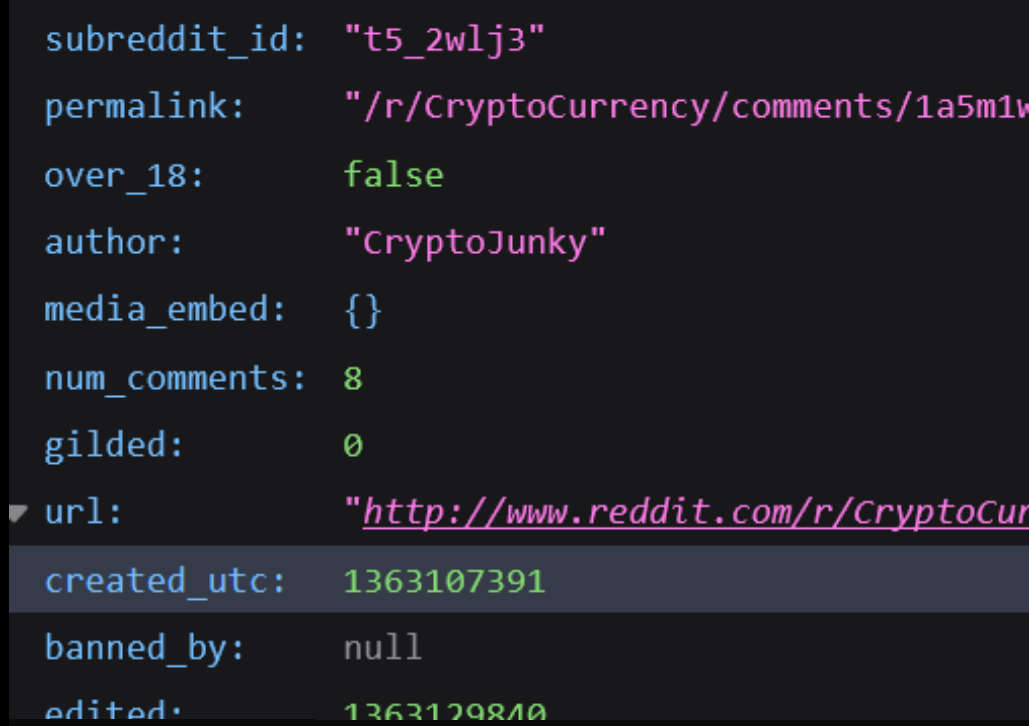
Pre-elaborazione

- Filtraggio di post e commenti rilevanti
- Eliminazione dei campi non necessari
- Salvataggio in formato JSON

Oltre ai dati di Reddit sono stati usati:

- Dati sulle ricerche da Google Trends
- Prezzi delle criptovalute da Binance

Sono stati poi uniti in vari file csv e utilizzati per svolgere varie analisi



	A	B	C	D
1	date	mean_comm	var_comme	std_comme
2	01/04/2022	2,98203E+14	1,97579E+14	1,40563E+15
3	02/04/2022	3,99519E+15	7,25857E+14	2,69417E+15
4	03/04/2022	5,36904E+15	8,16574E+14	2,85758E+15
5	04/04/2022	4,54461E+15	5,60366E+14	2,3672E+15
6	05/04/2022	3,67667E+15	5,04403E+14	2,24589E+15
7	06/04/2022	4,152E+15	9,71812E+14	3,11739E+15
8	07/04/2022	4,86105E+14	8,1851E+14	2,86096E+14
9	08/04/2022	3,40945E+15	4,61031E+14	2,14716E+15
10	09/04/2022	4,9841E+15	6,94213E+13	2,63479E+15
11	10/04/2022	3,70203E+15	4,43261E+13	2,10538E+14
12	11/04/2022	6,71615E+15	1,65742E+15	4,07115E+15
13	12/04/2022	4,23375E+15	9,09875E+14	3,01641E+15

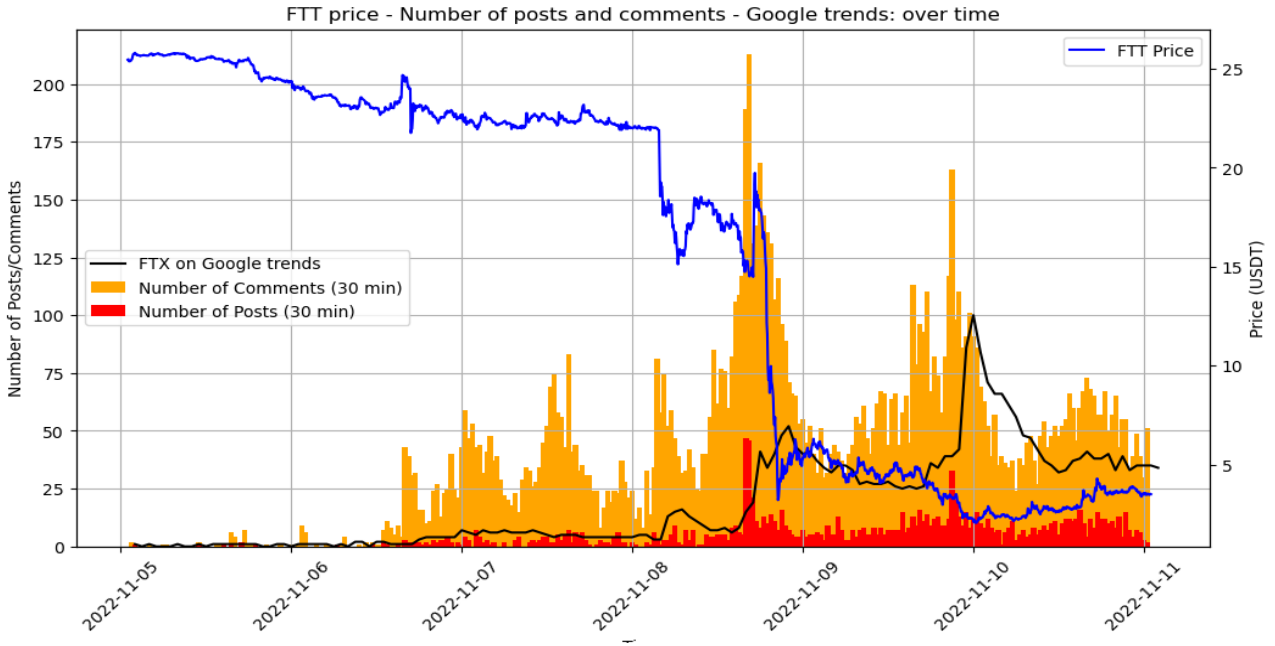
ANALISI SVOLTE

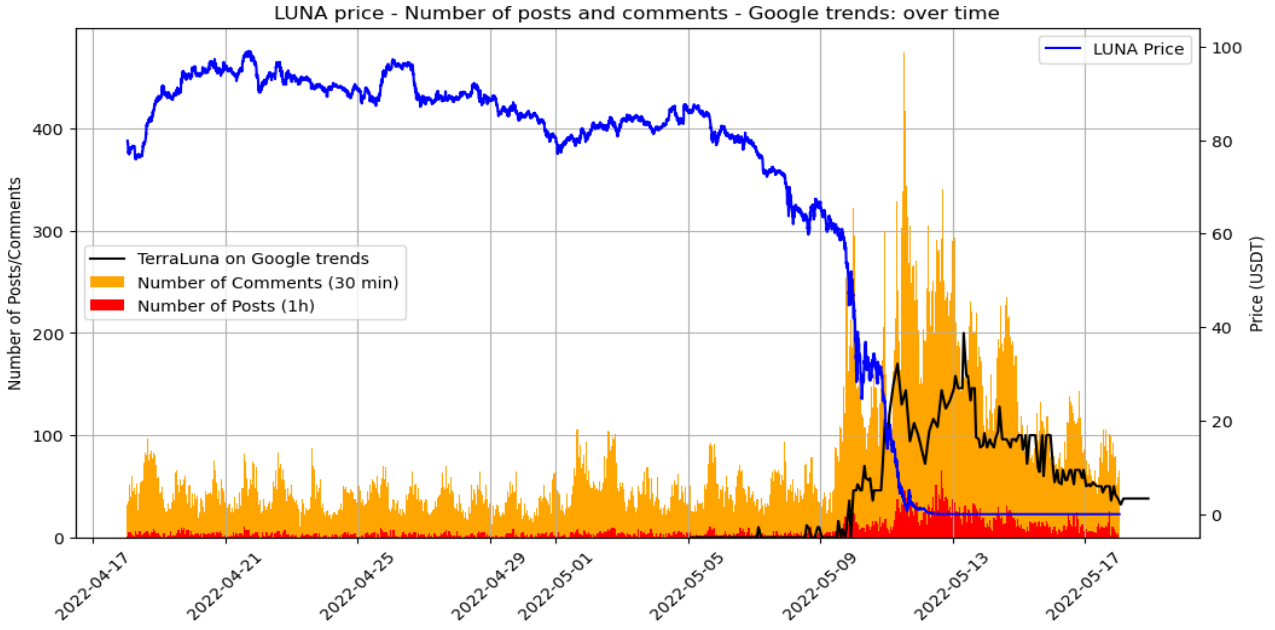
Quantitativa
- Grafici del numero di commenti e post
- Ricerca di un aumento significativo pre crollo

Sentiment:
- Utilizzo di Vader per classificare il sentiment
- Confronto con un periodo positivo per vedere se il sentiment era peggiore in concomitanza dei crolli

Google Trends
- Confronto delle ricerche con le discussioni sui social
- Per vedere se ci fosse una crescita nelle ricerche

Modellazione e Previsione
- Tentare di prevedere il prezzo utilizzando i vari dati estratti ed elaborati

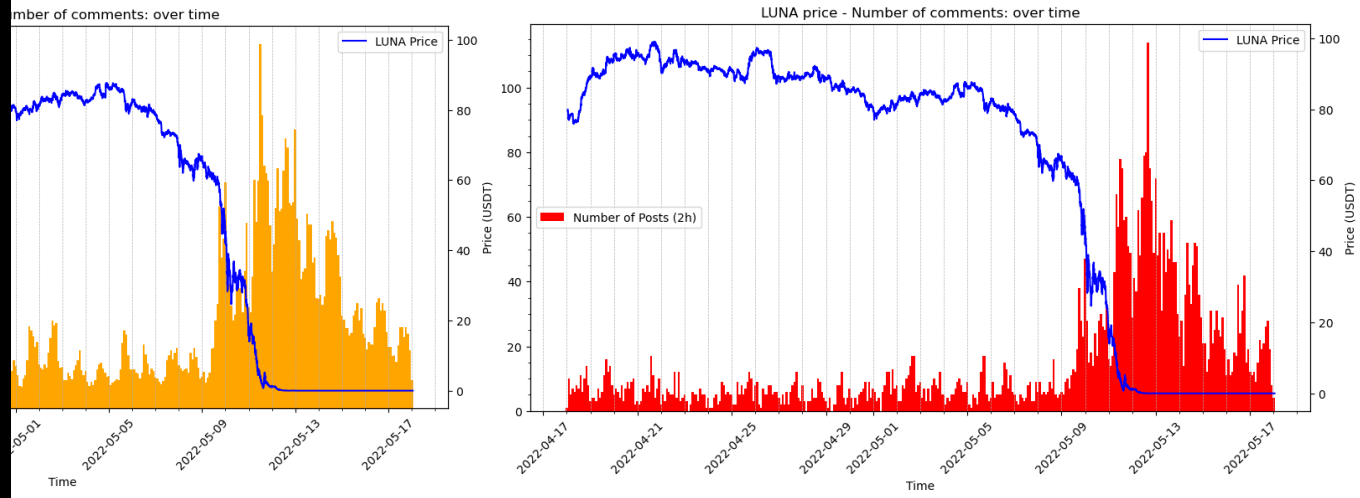


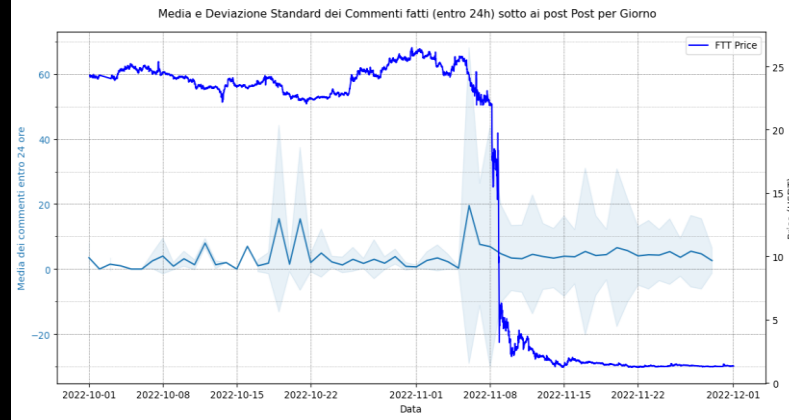


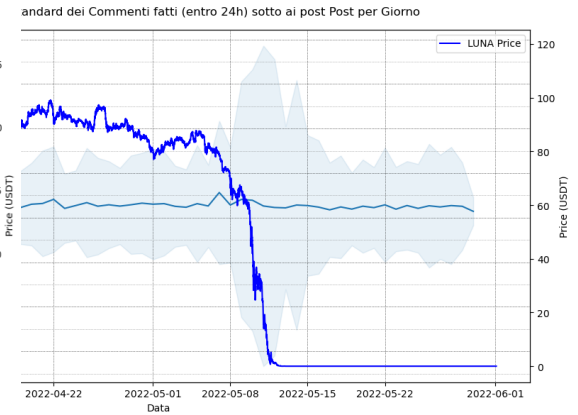
RISULTATI

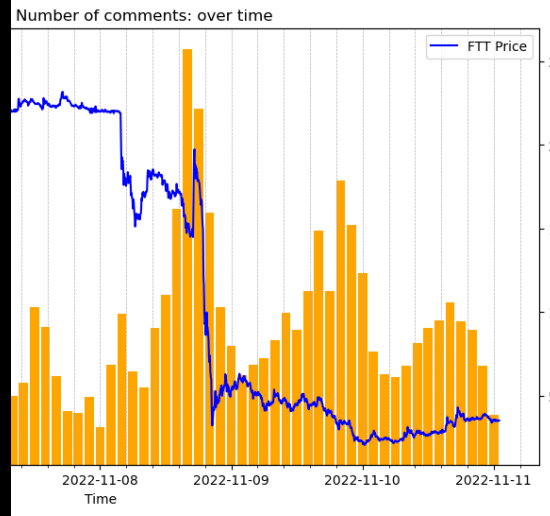
Analisi quantitativa

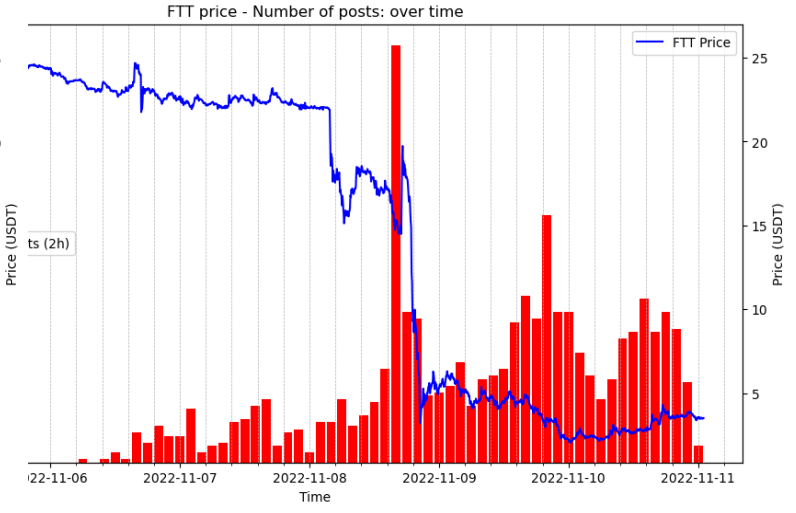
- Terra: i post e i commenti crescono di numero solo in concomitanza del crollo
- Ftx: aumento di post e commenti già prima del crollo
- Media e deviazione standard dei commenti fatti entro 24 ore dalla pubblicazione di post:
  - Terra: la deviazione standard aumenta col crollo, media circa stabile
  - Ftx: aumentano di entrambe già prima dell'evento







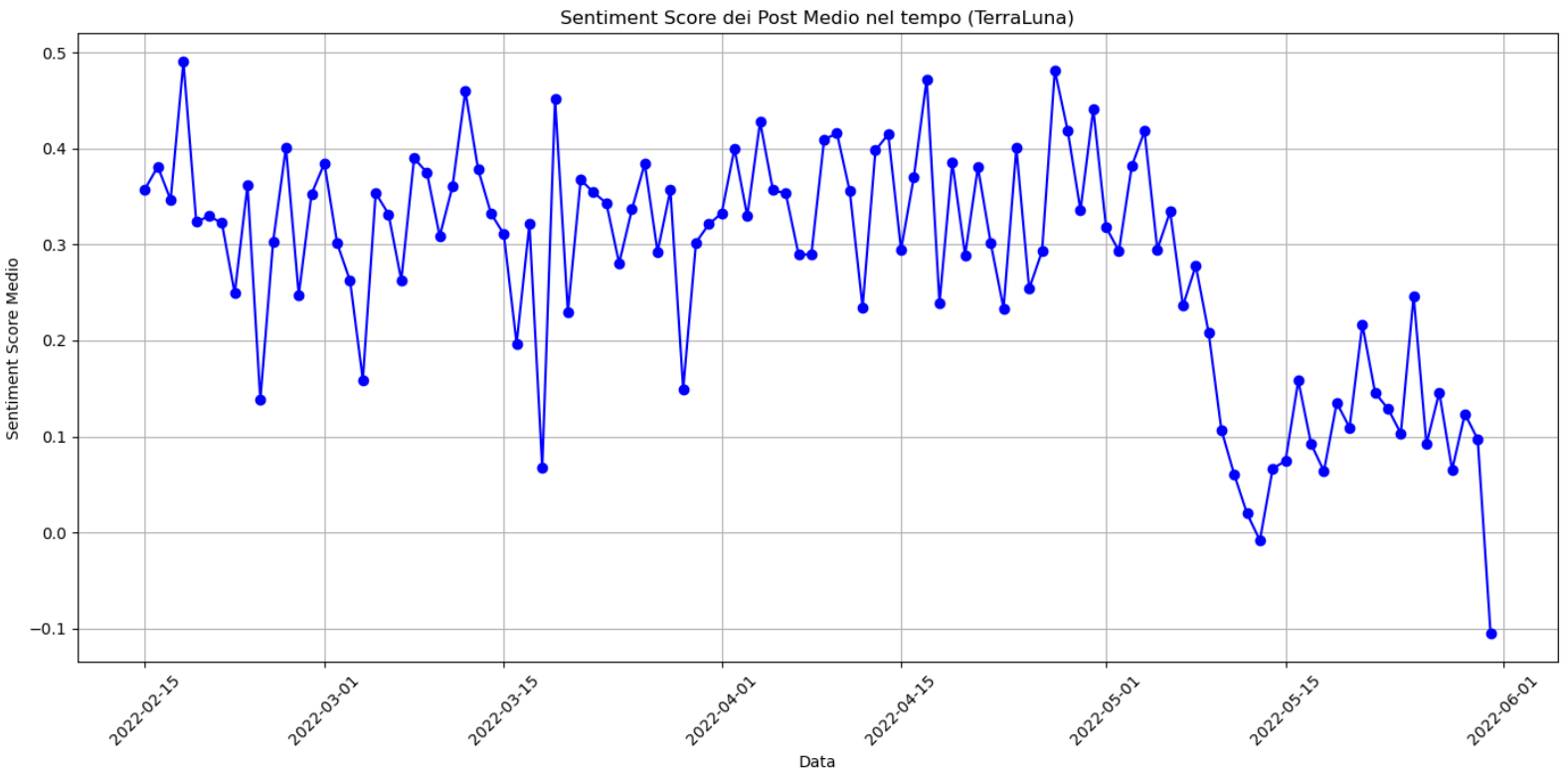


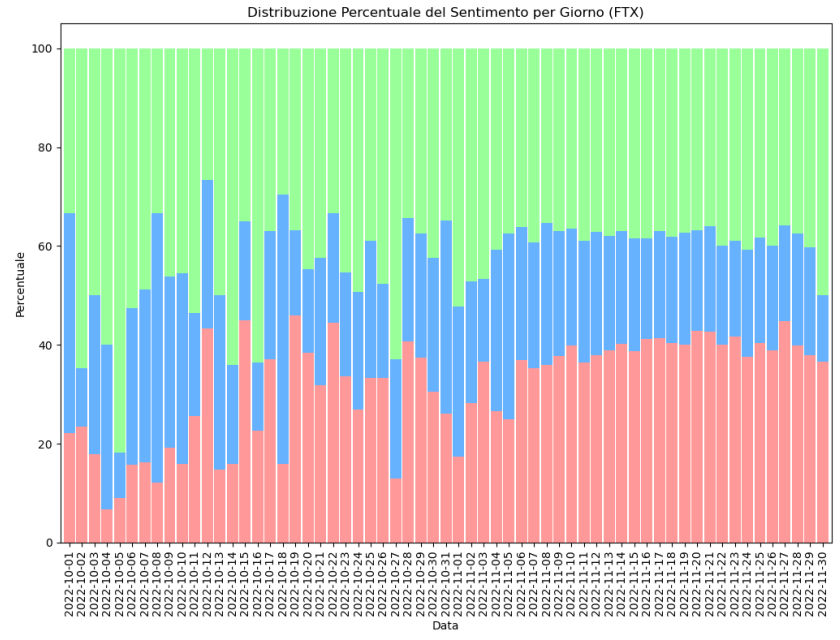


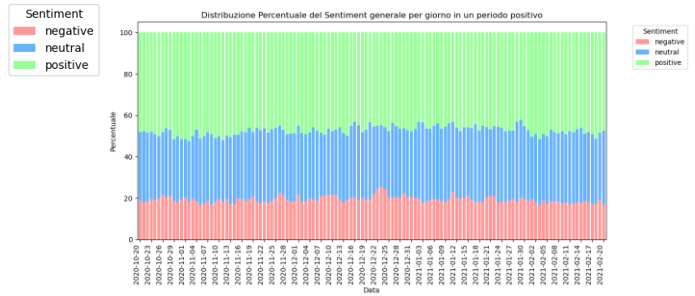
Sentiment

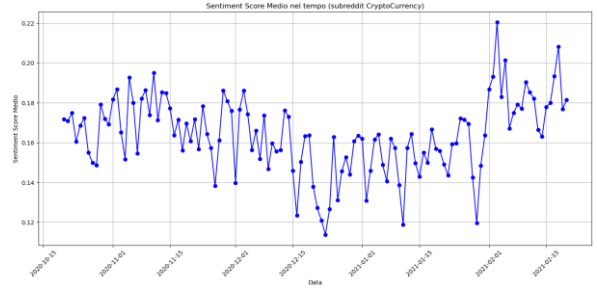
- Score di Vader: tende a 0 in prossimità del crollo (sui 0,16 di media in un periodo positivo)
- Distribuzione percentuale del sentiment: i post e commenti negativi stabili al 40% in prossimità del crollo (< 20% in un periodo positivo)

In basso a destra due grafici del sentiment in un periodo positivo di mercato







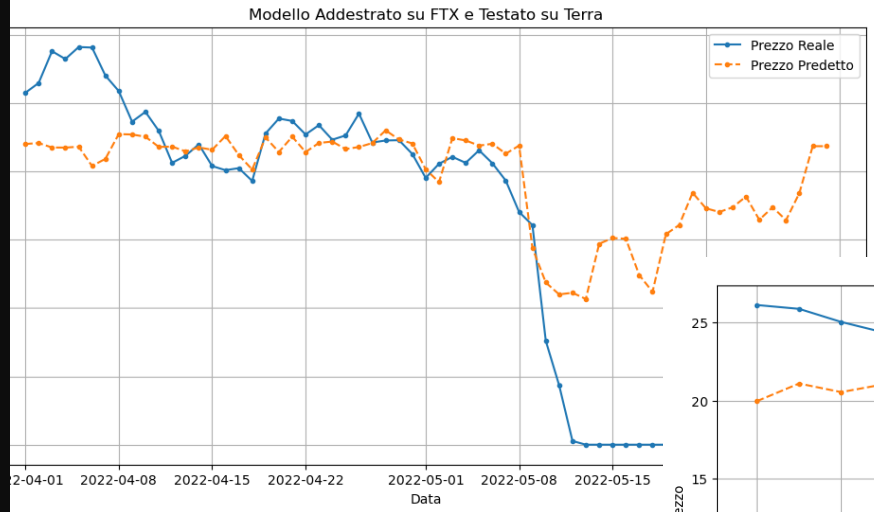


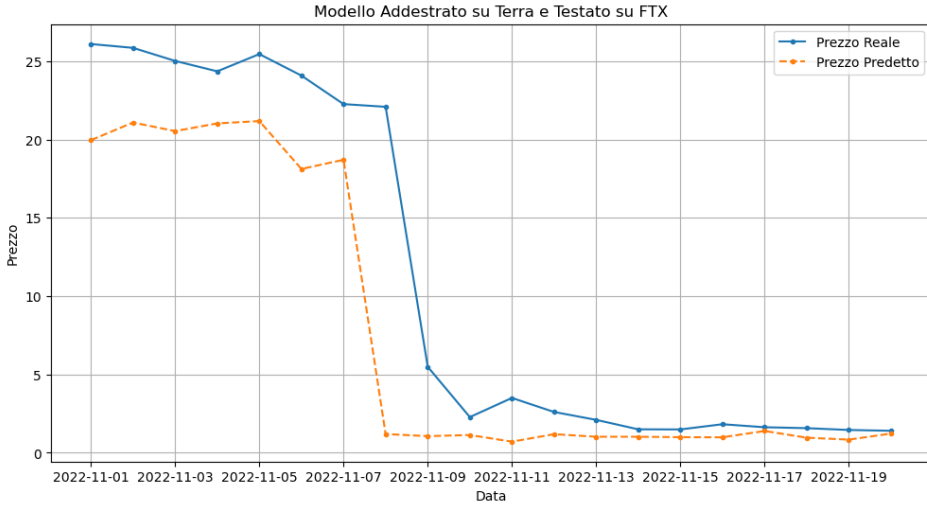
Modellazione e Previsione

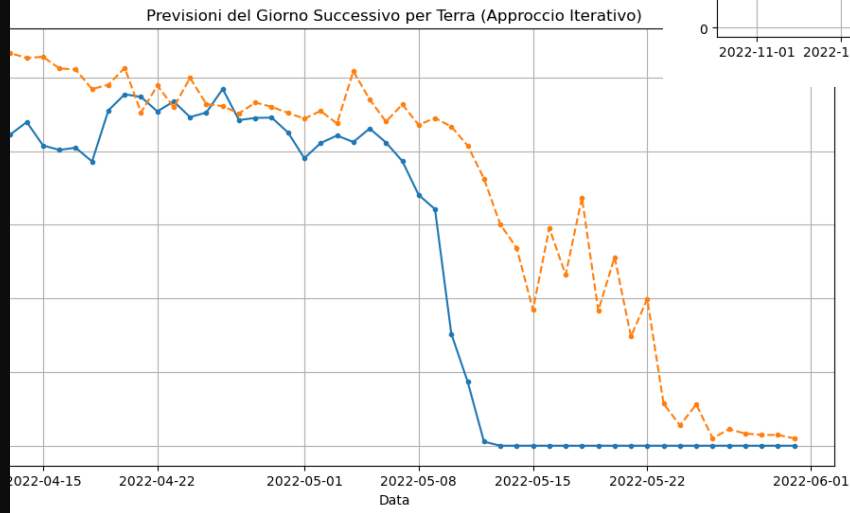
- Addestramento del Modello:
  - Modello di regressione Random Forest con scikit-learn
  - Valutazione con MSE e R^2

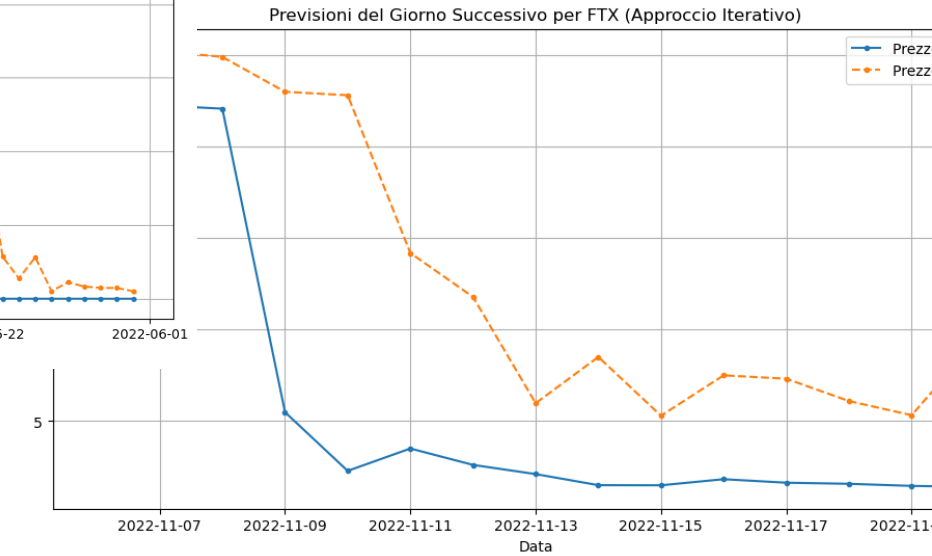
Approcci di Training e Test

1. Training e test sullo stesso dataset:
   - 80% training, 20% test
   - `train_test_split`
   - Terra: MSE=106, R^2=0,94
   - Ftx: MSE=15, R^2=0,89
2. Training e test su dataset separati:
   - Grid Search per ottimizzazione
   - Terra: MSE=1524, R^2=0,2
   - Ftx: MSE=30, R^2=0,75
3. Training iterativo:
   - Addestramento giornaliero
   - Previsione giorno successivo
   - Terra: MSE=787, R^2=0,55
   - Ftx: MSE=103, R^2=0,5









CONCLUSIONE

- L'analisi dei dati dei social media può fornire alcuni piccoli segnali di allarme tempestivi
- Il modello di machine learning ha mostrato risultati promettenti, ma richiede ulteriori miglioramenti

Miglioramenti futuri:

- Una maggiore base di dati
- Ulteriori ricerche per migliorare accuratezza