POLITECNICO
MILANO 1863

Utilizzo dei dati dei social media per l'analisi del sentiment e la previsione dei crolli delle criptovalute: un caso di studio su Terra e FTX

Progetto di ingegneria informatica

A.A. 2023/2024

Autore:

Giacomo Falcone
(10804110)

giacomo.falcone@mail.polimi.it

Contenuti

1 Introduzione

Negli ultimi anni, le criptovalute hanno acquisito un'importanza crescente nel panorama finanziario globale. Originariamente concepite come forme alternative di valuta digitale, Bitcoin e altre criptovalute sono rapidamente diventate strumenti di investimento e speculazione largamente riconosciuti. La continua espansione del mercato delle criptovalute ha attirato l'interesse di investitori privati, istituzioni finanziarie e regolatori, riflettendo la crescente rilevanza di questi asset nel mondo contemporaneo.

Uno degli aspetti più distintivi e affascinanti delle criptovalute è la loro estrema volatilità. A differenza delle valute tradizionali e di altri asset finanziari, i prezzi delle criptovalute possono subire oscillazioni drastiche in periodi di tempo molto brevi. Questa volatilità rende il mercato delle criptovalute altamente imprevedibile e suscettibile a una serie di fattori, tra cui notizie di regolamentazioni, attacchi hacker a piattaforme di scambio e sentiment degli investitori.

Nel contesto di questo scenario dinamico e spesso imprevedibile, il presente progetto si propone di analizzare i dati dei social media per prevedere l'andamento delle criptovalute. In particolare, l'analisi si è concentrata su post e commenti nel subreddit CryptoCurrency [10] durante periodi specifici, al fine di identificare segnali di allarme che possano indicare un possibile crollo imminente di una criptovaluta.

Sono stati esaminati due casi di crollo significativi: quello di FTX e quello di Terraform Labs. L'obiettivo principale del progetto è stato capire se è possibile rilevare segnali precoci di un crollo attraverso l'analisi delle discussioni online. Le criptovalute, essendo estremamente volatili, possono crescere rapidamente, ma altrettanto rapidamente possono crollare. Pertanto, uno strumento che possa fornire avvisi tempestivi è fondamentale non solo per i trader attivi, ma anche, e forse soprattutto, per coloro che "holdano" le criptovalute, ovvero le acquistano e le conservano a lungo termine senza monitorare costantemente il mercato.

Per raggiungere questo obiettivo, sono stati utilizzati i dati del subreddit CryptoCurrency [10], selezionando le date intorno ai due eventi di interesse e filtrando i commenti e i post contenenti parole chiave correlate agli eventi stessi. Sono stati analizzati il numero di post e commenti nel tempo e integrati questi dati con le tendenze di ricerca di Google Trends per verificare se le ricerche relative agli eventi aumentassero. Un'ulteriore analisi è stata condotta utilizzando Vader per determinare il sentiment dei commenti e dei post, per valutare se si notasse un aumento dei commenti negativi come segno premonitore di un crollo imminente.

Infine, questi dati sono stati utilizzati per addestrare un modello di machine learning con l'obiettivo di prevedere i prezzi delle criptovalute. Il modello è stato addestrato su un evento specifico e testato su un altro, ottenendo risultati promettenti sebbene non perfetti. Questo approccio dimostra il potenziale dell'analisi dei dati dei social media per la previsione dei movimenti di mercato delle criptovalute, ponendo le basi per lo sviluppo di un possibile strumento prezioso per gli investitori.

2 Related Work

Negli ultimi anni, diversi studi hanno esplorato metodi innovativi per la previsione dei prezzi delle criptovalute, combinando analisi delle serie temporali con l'analisi del sentiment proveniente dai social media. Di seguito è presentata una panoramica dei lavori rilevanti esaminati per questo progetto:

Uno studio significativo esplora l'uso combinato dei dati storici dei prezzi e dei punteggi di sentiment derivati dai social media per prevedere il prezzo del Bitcoin. Gli autori utilizzano modelli avanzati di previsione delle serie temporali arricchiti con dati di sentiment per migliorare l'accuratezza delle previsioni. In particolare, viene utilizzato un modello di regressione lineare combinato con tecniche di analisi del sentiment per migliorare la previsione dei prezzi. I risultati mostrano un miglioramento significativo rispetto ai modelli che utilizzano solo dati storici [1].

Un altro articolo esamina l'uso dell'analisi del sentiment su Twitter per la previsione dei prezzi delle criptovalute. Gli autori implementano un modello che integra i dati di sentiment dei tweet con modelli di machine learning come la regressione lineare e le reti neurali per prevedere i movimenti dei prezzi delle criptovalute. L'analisi mostra che il sentiment sui social media può fornire informazioni utili per le previsioni dei prezzi, con modelli che includono dati di sentiment che mostrano una maggiore accuratezza rispetto ai modelli che non lo fanno [2].

In uno studio innovativo, gli autori introducono un approccio che utilizza modelli BERT per l'analisi del sentiment e una tecnica di supervisione debole per affrontare la mancanza di etichette nei dati testuali. Utilizzando il modello BERT, riescono a estrarre caratteristiche testuali avanzate che migliorano la capacità predittiva del sentiment. La supervisione debole consente di utilizzare un grande volume di dati non etichettati, migliorando ulteriormente la robustezza del modello. I risultati mostrano che l'uso di etichette deboli migliora il valore predittivo delle caratteristiche testuali, aumentando l'accuratezza delle previsioni dei rendimenti delle criptovalute [3].

Infine, un lavoro esplora vari modelli di machine learning per la previsione dei prezzi delle criptovalute, integrando l'analisi del sentiment proveniente dai social media. Gli autori confrontano l'efficacia di diversi algoritmi di machine learning, tra cui regressione lineare, support vector machines e reti neurali, concludendo che l'integrazione dei dati di sentiment può migliorare significativamente le performance dei modelli di previsione. L'analisi dettagliata mostra come diversi approcci di machine learning rispondono all'inclusione di dati di sentiment e quali modelli sono più adatti per questo tipo di previsione [4].

3 Metodo

Segue una descrizione generale del metodo adottato per questo progetto. Le informazioni sugli esperimenti verranno fornite nella sezione successiva (4).

3.1 Raccolta dei Dati

I dati utilizzati nel progetto sono stati estratti dal subreddit `CryptoCurrency` [10] utilizzando i dump di Pushshift [8], accessibili tramite Academic Torrents [9]. I dati sono stati scaricati utilizzando un server `qBittorrent`, ottenendo i dati in formato di cartelle compresse `.zst`. Successivamente, sono stati utilizzati script Python per decomprimere e filtrare i dati, salvando solo i post e i commenti di circa tre mesi intorno agli eventi di interesse (FTX e Terraform Labs) in due file JSON separati.

3.2 Pre-elaborazione dei Dati

Dai file JSON, sono stati ulteriormente estratti i post e i commenti contenenti parole chiave collegate agli eventi di interesse, eliminando i campi non necessari e salvandoli in altri due file JSON. Questo ha permesso di concentrarsi solo sui dati rilevanti per l'analisi.

3.3 Analisi svolte

3.3.1 Analisi Quantitativa

Sono stati tracciati grafici del numero di commenti e post nel tempo per identificare pattern significativi e determinare i periodi di maggiore interesse. Questa analisi preliminare ha evidenziato aumenti di attività intorno ai crolli delle criptovalute.

3.3.2 Analisi del Sentiment

Utilizzando lo strumento Vader [11], i post e i commenti sono stati classificati in sentiment positivo, negativo e neutro. I punteggi del sentiment sono stati confrontati con un periodo positivo del mercato.

3.3.3 Integrazione con Google Trends

I dati di Google Trends [12] sono stati utilizzati per confrontare le ricerche relative agli eventi con le discussioni sul subreddit, analizzando la correlazione tra le attività di ricerca e i volumi di post e commenti.

3.3.4 Media e Deviazione Standard dei Commenti per Post

Utilizzando Pandas [5], i DataFrame dei post e dei commenti sono stati uniti tramite l'ID, calcolando la media e la deviazione standard giornaliera dei commenti per ciascun post entro 24 ore dalla sua pubblicazione. Questa analisi è stata utilizzata per identificare anomalie e aumenti significativi di attività.

3.4 Modellazione e Previsione

Tutti i dati rilevanti sono stati uniti in un unico DataFrame e salvati in formato CSV. Utilizzando scikit-learn [7], è stato addestrato un modello di regressione Random Forest, normalizzando le feature prima dell'addestramento. Prima dell'addestramento, anche il target (prezzo) è stato normalizzato dividendo ogni valore per il massimo osservato all'interno di ciascun dataset, per garantire che i valori fossero su scale comparabili. La performance del modello è stata valutata confrontando le previsioni con i valori reali, utilizzando la Mean Squared Error (MSE) e il coefficiente di determinazione (R^2) [27].

Per la fase di training e test, sono stati adottati tre approcci distinti:

1. **Training e test sullo stesso dataset:** Il dataset è stato suddiviso in due parti, utilizzando una percentuale per l'addestramento e una per il test (80% training e 20% test). Questa metodologia ha permesso di valutare le prestazioni del modello su dati mai visti ma appartenenti allo stesso dataset.

- I dati sono stati divisi in training e test set utilizzando train_test_split con una percentuale di test del 20% e un random state di 42.
- È stato creato un modello di regressione Random Forest e addestrato sui dati di training.
- Sono state fatte previsioni sul test set e il modello è stato valutato utilizzando la Mean Squared Error (MSE).

2. **Training e test su dataset separati**: Il modello è stato addestrato su un intero dataset (ad esempio, FTX) e poi testato su un altro dataset completo (ad esempio, Terra). Questo metodo consente di valutare la capacità del modello di generalizzare su dati completamente nuovi e differenti.
   - È stata definita una griglia di parametri per ottimizzare il modello Random Forest utilizzando Grid Search [29].
   - Il modello è stato addestrato sui dati normalizzati del dataset FTX e poi testato sui dati normalizzati del dataset Terra (e viceversa).
   - Le previsioni sono state denormalizzate per confrontarle con i prezzi reali e il modello è stato valutato utilizzando la MSE.
3. **Training iterativo**: Per ogni giorno, il modello è stato addestrato utilizzando tutti i dati disponibili fino a quel giorno. La previsione è stata poi effettuata per il giorno successivo. Questo processo è stato ripetuto iterativamente per tutto il dataset, partendo da un periodo iniziale di alcuni giorni di dati storici.
   - È stata impostata una dimensione iniziale di alcuni giorni per il training set.
   - Per ogni giorno successivo, il modello è stato addestrato con i dati disponibili fino a quel giorno e il prezzo del giorno successivo è stato previsto.
   - Le previsioni sono state denormalizzate per confrontarle con i prezzi reali e il modello è stato valutato utilizzando la MSE.

4 Esperimenti: dati e risultati

4.1 Raccolta e Pre-elaborazione dei Dati

I dati sono stati raccolti dal subreddit `CryptoCurrency` utilizzando i dump di Pushshift [8]. Utilizzando un client qBittorrent, i dataset sono stati scaricati da Academic Torrents in formato compresso `.zst`. Successivamente, degli script Python sono stati impiegati per decomprimere e filtrare i dati, selezionando solo i post e i commenti pertinenti agli eventi di interesse (FTX e Terraform Labs). Per ciascun evento, è stato considerato un intervallo di circa tre mesi. Questi dati sono stati quindi salvati in formato JSON per le successive analisi.

Inoltre, i dati sono stati ulteriormente filtrati mantenendo solo i campi rilevanti per l'analisi, come il titolo, il corpo dei commenti, lo score, ecc. Sono stati salvati solo quei post e commenti che contenevano le parole chiave rilevanti per i due eventi:

- Keywords per Terraform Labs: 'Terra', 'TerraUSD', 'Luna', 'Do Kwon', 'Terraform Labs', 'UST Anchor Protocol', 'UST depeg', 'Kwon Do-Hyung', ('terra', 'terrausd', 'luna', 'terraform', 'terraformlabs', 'ust', 'do kwon', 'kwon do-hyung')
- Keywords per FTX: 'FTX', 'FTT', 'Alameda Research', 'Sam Bankman-Fried', 'sbf', 'FTX token', ('ftx', 'ftt', 'alameda', 'sbf', 'bankman-fried')

4.2 Analisi Preliminare

Sono stati creati grafici per visualizzare il numero di post e commenti nel tempo, evidenziando picchi di attività intorno ai crolli delle criptovalute. Questi grafici hanno permesso di identificare periodi di maggiore interesse, consentendo di ridurre ulteriormente l'intervallo temporale da analizzare. In questo modo, è stato possibile focalizzare l'analisi su periodi più dettagliati, migliorando la granularità temporale dello studio.

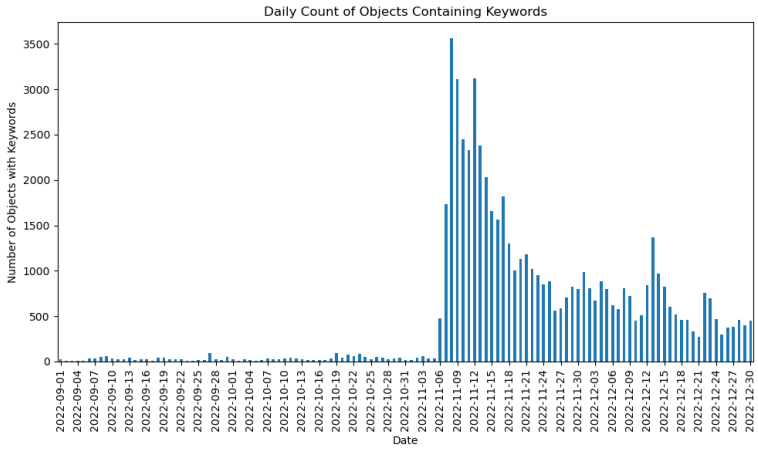

Figura 1: Numero di post al giorno su FTX

4.2.1 Prezzo e numero di post e commenti

Per analizzare la correlazione tra il prezzo delle criptovalute e l'attività sui social media, sono stati creati grafici che combinano il prezzo delle criptovalute con il numero di post e commenti. I valori sul prezzo delle criptovalute sono stati raccolti dal sito di Binance [17] in formato csv. I dati sono stati divisi in intervalli di 2 ore per mantenere un'elevata granularità, assicurando al contempo un volume sufficiente di commenti e post nei vari istogrammi.

Dai grafici allegati, si possono notare alcune tendenze significative:

Caso FTX (Figura 3 e 5): il numero di post e commenti su FTX mostra un aumento significativo già tempo prima del crollo. Si nota infatti come prima del 6 novembre 2022 il numero di post e commenti sull'argomento sia quasi nullo, mentre il 7 novembre avviene un notevole aumento. Se si va a cercare cosa accadde quel giorno si trova che il 6 e 7 novembre 2022 sono stati giorni cruciali per il crollo di FTX.

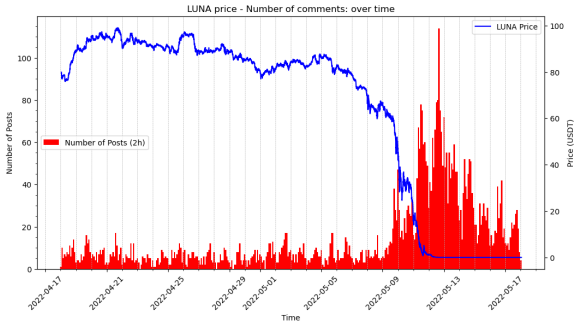

Figura 2: Numero di post su Terra

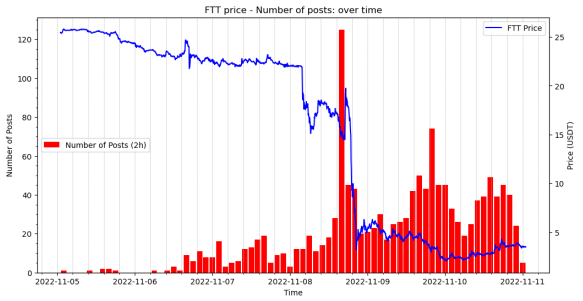

Figura 3: Numero di post su FTX

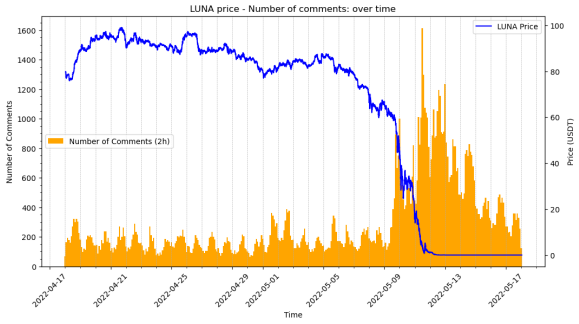

Figura 4: Numero di commenti su Terra

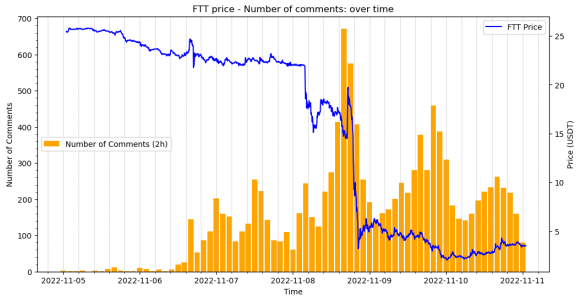

Figura 5: Numero di commenti su FTX

- 6 novembre 2022: Changpeng Zhao (CZ), CEO di Binance, ha annunciato su Twitter che Binance avrebbe venduto tutte le sue partecipazioni nel token FTT di FTX, a causa di recenti rivelazioni sulle condizioni finanziarie di FTX e del suo legame con Alameda Research [13, 14].

- 7 novembre 2022: Sam Bankman-Fried (SBF), CEO di FTX, ha cercato di rassicurare gli investitori affermando che "gli asset sono a posto" e che FTX era in buona salute finanziaria. Tuttavia, questa affermazione non è riuscita a placare le preoccupazioni degli investitori, e le speculazioni sulla solvibilità di FTX hanno continuato a crescere [15, 16].

Questi eventi hanno portato a un massiccio ritiro di fondi da FTX, creando una crisi di liquidità che ha contribuito al crollo del prezzo del token FTT e alla successiva dichiarazione di bancarotta di FTX.

In corrispondenza del crollo la quantità di post e commenti aumenta vertiginosamente restando molto alta anche successivamente al crollo.

Caso Terraform Labs (Figura 2 e 4): A differenza di FTX qui si osserva un notevole aumento del numero di post e commenti solo intorno al periodo del crollo di LUNA. Se si va a guardare più nel dettaglio la cronologia degli eventi su questo fatto, si trova che:

- 7 maggio 2022: UST [25] inizia a perdere il suo peg [24] a causa del ritiro di liquidità da Curve [23].

- 8 maggio 2022: LFG (Luna Foundation Guard) [26] annuncia un'iniezione di 1,5 miliardi di dollari per sostenere UST, ma la fiducia è già compromessa.

- 9 maggio 2022: Le vendite di UST aumentano, e il prezzo di LUNA inizia a crollare significativamente.
- 10-12 maggio 2022: La fornitura di LUNA cresce esponenzialmente, mentre il prezzo crolla quasi a zero [18, 19].

Quindi a differenza del caso di FTX, dove c'erano segnali e discussioni che precedevano il crollo, il crollo di Terra è stato estremamente rapido e senza precedenti avvertimenti pubblici significativi. La complessità del sistema e la rapidità degli eventi hanno impedito una risposta tempestiva da parte degli investitori meno informati. Solo dopo il crollo è emerso che alcune parti del sistema erano insostenibili e che le riserve di LFG non erano sufficienti per sostenere l'ecosistema [19, 20].

4.3 Analisi del Sentiment

Per l'analisi del sentiment è stato utilizzato VADER (Valence Aware Dictionary and sEntiment Reasoner) [11], un modello di analisi del sentiment basato su regole e lessico progettato per i contesti dei social media. VADER assegna un punteggio di sentiment a ogni testo, classificandolo come positivo, negativo o neutrale in base a soglie predefinite (score ≥ 0.05: positivo, score ≤ 0.05: negativo, altrimenti neutrale).

L'analisi è stata condotta su un periodo di circa due mesi per ciascun caso (FTX e Terra) per osservare eventuali cambiamenti nel sentiment rispetto al passato. Sono stati generati tre tipi di grafici per ciascun caso, oltre a un confronto con un periodo di mercato positivo (la crescita di Bitcoin a fine 2020 e inizio 2021):

- Distribuzione Percentuale del Sentiment per Giorno: Mostra la percentuale di post classificati come positivi, neutri o negativi ogni giorno.
- Sentiment Score Medio Giornaliero dei Post: Rappresenta la media giornaliera del punteggio di sentiment assegnato da VADER ai post.
- Sentiment Score Medio Giornaliero dei Commenti: Simile al precedente, ma applicato ai commenti.

4.3.1 Distribuzione Percentuale del Sentiment per Giorno

Caso Terra (Figura 6): Nel caso di Terra, non si osservano variazioni significative nel tempo. Tuttavia, il 9 maggio il sentiment positivo scende sotto il 50%, dopo essere rimasto sopra tale soglia per oltre due mesi. Questo cambiamento coincide con l'inizio del crollo dell'ecosistema Terra. Sebbene il prezzo non fosse ancora crollato drasticamente (cosa che avverrà il giorno successivo), questo rappresenta un piccolo segnale di allarme anticipato, anche se non particolarmente significativo o sostanzioso.

Caso FTX (Figura 7): Per FTX, i dati sui post e sui commenti fino al 6 novembre sono pochi e dunque il sentiment oscilla molto. Dal 6 novembre, quando l'argomento inizia a essere discusso, la percentuale di sentiment negativo si assesta intorno al 40%. Confrontando questo dato con il periodo di mercato positivo, in cui i post e commenti negativi erano intorno al 20%, si nota una significativa differenza.

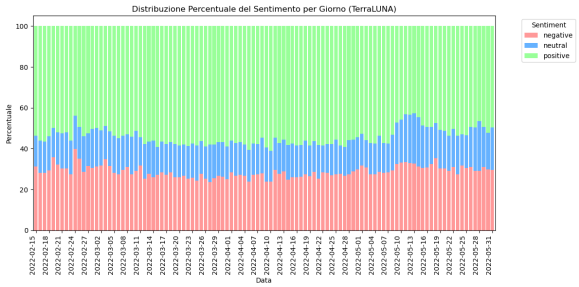

Figura 6: Sentiment sul caso Terra

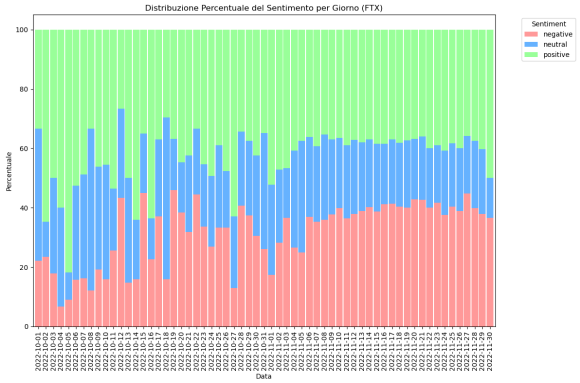

Figura 7: Sentiment sul caso FTX

4.3.2 Sentiment Score Medio Giornaliero

Caso Terra (Figura 8): Il sentiment score medio giornaliero per Terra mostra un calo iniziato il 9 maggio 2022. Sebbene la media di quel giorno fosse inferiore rispetto al mese precedente, non era significativamente più bassa. Dal 10 maggio, il sentiment scende notevolmente, assestandosi attorno allo zero.

Caso FTX (Figura 9): Anche per FTX, lo score tende a zero in concomitanza del crollo. Nei giorni precedenti, lo score oscilla molto e si stabilizza solo dopo che l'argomento diventa più discusso. Questo indica che l'analisi del sentiment è meno utile quando i dati sono scarsi. Al contrario, nel caso di Terra, l'argomento era già discusso nei mesi precedenti, mostrando un andamento più omogeneo e sensato.

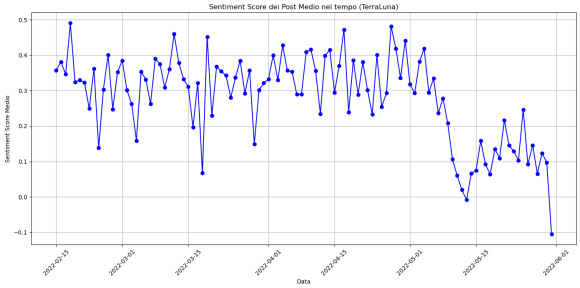

Figura 8: Sentiment score di Terra

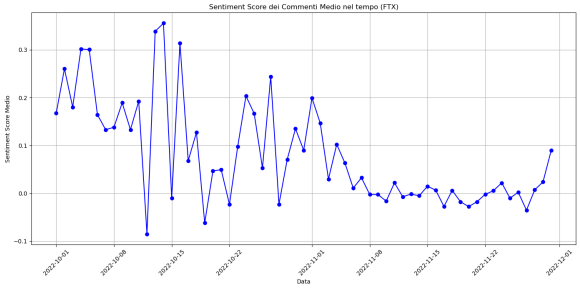

Figura 9: Sentiment score di FTX

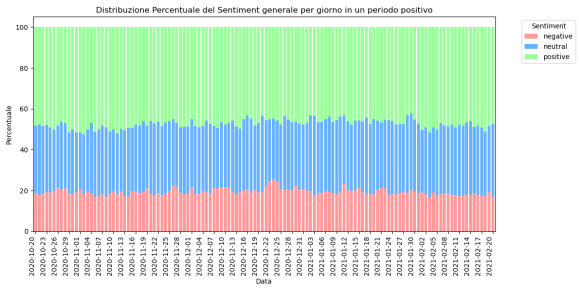

Figura 10: Sentiment nel periodo positivo

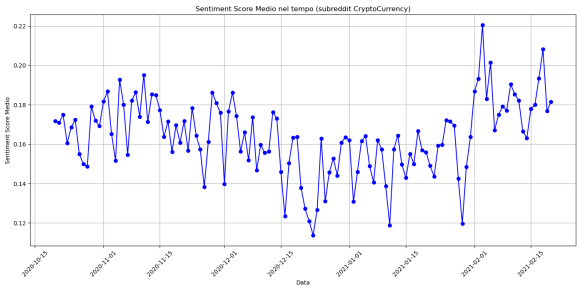

Figura 11: Score nel periodo positivo

4.3.3 Confronto con Periodo di Mercato Positivo

Confrontando i dati dei due casi con il periodo di mercato positivo, emerge un'informazione importante: il sentiment score si assesta intorno allo zero per entrambi i casi di crollo, mentre durante il periodo positivo del mercato il valore è circa 0,15. La percentuale di sentiment negativo in questo periodo è intorno al 20%, minore rispetto ai due casi analizzati.

4.4 Integrazione con Google Trends

I dati di Google Trends [12] sono stati utilizzati per confrontare le ricerche relative agli eventi di interesse con l'attività sui social media e il prezzo delle due criptovalute.

Per analizzare il coefficiente di interesse, ovvero il numero di ricerche su Google per specifici termini di ricerca, sono stati selezionati vari termini rilevanti per ciascun evento. Questi termini includono parole chiave associate agli eventi di crollo di Terra e FTX. Per ciascun termine, è stato estratto l'indice di interesse da Google Trends. In seguito, è stata calcolata una media dei valori ottenuti per ottenere un quadro complessivo delle ricerche correlate agli eventi.

4.4.1 Procedura di Raccolta dei Dati

1. **Selezione dei Termini di Ricerca**: I termini di ricerca sono stati scelti in base alla loro rilevanza per ciascun evento (simili alle parole chiave usate per il filtraggio di commenti e post).

2. **Estrazione dei Dati**: Utilizzando Google Trends, sono stati estratti i dati di interesse per ciascun termine di ricerca. Questi dati rappresentano l'indice di interesse su una scala da 0 a 100, dove 100 indica il picco di popolarità del termine. Ciascun termine è stato cercato insieme allo stesso termine di paragone ("matita") in modo da avere un riferimento costante, rendendo i risultati più comparabili tra loro. Il termine "matita" è stato selezionato perché è un termine comune, cercato in modo relativamente costante nel tempo, senza picchi di popolarità troppo elevati o troppo bassi.

3. **Calcolo della Media**: Per ottenere un indice di interesse più robusto e rappresentativo, è stata calcolata la media dei valori di interesse per tutti i termini di ricerca associati a ciascun evento.

4.4.2 Risultati

I risultati dell'analisi con Google Trends mostrano un pattern simile a quello osservato nei grafici del numero di commenti e post sui social media. Questo suggerisce che vi è una correlazione tra l'interesse di ricerca su Google e l'attività sui social media nei periodi di crisi.

Caso Terra (Figura 12): L'indice di interesse su Google Trends per i termini relativi a Terra mostra un picco significativo intorno al 9 maggio 2022, in concomitanza con l'inizio del crollo del prezzo di LUNA. Questo riflette l'aumento delle ricerche su Google riguardanti Terra e i suoi componenti chiave, suggerendo che l'interesse pubblico è cresciuto rapidamente al diffondersi delle notizie sul crollo.

Caso FTX (Figura 13): Per FTX, si osserva un aumento dell'indice di interesse su Google Trends a partire dal 6 novembre 2022, giorno in cui sono emerse le notizie critiche sullo

stato finanziario di FTX. L'indice di interesse raggiunge un picco nei giorni immediatamente successivi, coincidente con il crollo del prezzo di FTT.

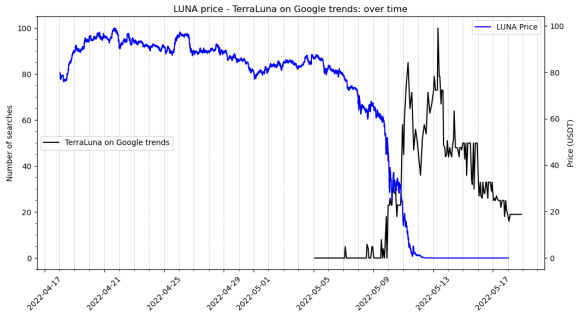

Figura 12: Ricerche Google su Terra

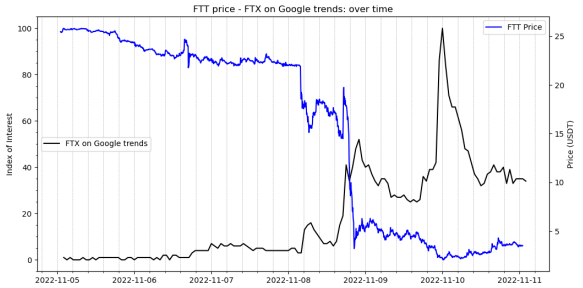

Figura 13: Ricerche Google su FTX

4.5 Andamento dei Commenti per Post

Sono state calcolate la media (μ) e la deviazione standard (σ) giornaliera dei commenti per post, rivelando pattern significativi di attività. Per ottenere questi risultati, i dati dei post e dei commenti sono stati uniti tramite l'ID dei post, seguendo questi passaggi:

1. **Unione dei Dati**: I file JSON contenenti i post e i commenti sono stati uniti utilizzando il campo `link_id` nei commenti, che corrisponde all'ID dei post senza il prefisso `t3_`.

2. **Filtraggio dei Dati**: Sono stati mantenuti solo i dati rilevanti per l'analisi, eliminando i campi non necessari.

3. **Conteggio dei Commenti**: Per ogni post, è stato calcolato il numero di commenti ricevuti entro 24 ore dalla pubblicazione del post. Questo conteggio è stato fatto utilizzando Pandas per creare un DataFrame che associa a ciascun post il numero di commenti ricevuti in diversi intervalli temporali.

4. **Calcolo delle Statistiche**: Sono stati calcolati la media e la deviazione standard giornaliera del numero di commenti per post.

Risultati: I grafici risultanti (Figure 14 e 15) mostrano la media e la deviazione standard dei commenti per post, oltre al prezzo delle criptovalute nel tempo. Questi grafici forniscono una visione chiara dell'attività sui social media in relazione alle fluttuazioni di prezzo delle criptovalute.

4.5.1 Caso Terra (Figura 14)

Nel caso di Terra, il grafico mostra che la media dei commenti per post non aumenta significativamente durante il periodo del crollo. Tuttavia, si osserva un aumento della deviazione standard, il che indica una maggiore variabilità nel numero di commenti per post. Questo suggerisce che ci sono stati pochi post con molti commenti e molti post con pochi commenti. Questo pattern è visibile in concomitanza con il crollo del prezzo di LUNA e nei giorni successivi, ma non fornisce segnali significativi di allarme anticipato.

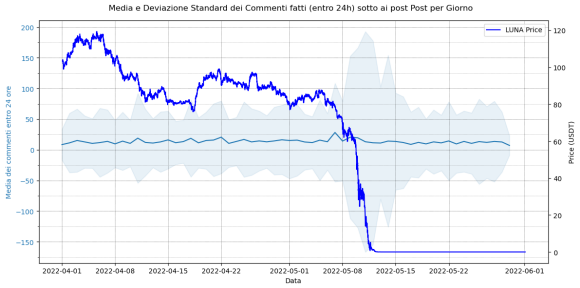

Figura 14: μ e σ sul caso Terra

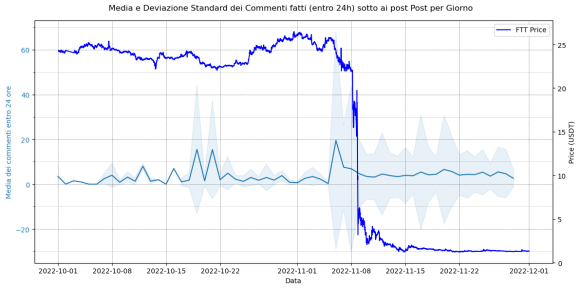

Figura 15: μ e σ sul caso FTX

4.5.2 Caso FTX (Figura 15)

Nel caso di FTX, il grafico mostra un aumento sia della media dei commenti per post che della deviazione standard a partire da metà ottobre 2022. Questo periodo coincide con diverse notizie critiche riguardanti FTX:

- 16 ottobre 2022: È stato rivelato che FTX era sotto indagine da parte della SEC (Securities and Exchange Commission) degli Stati Uniti per possibili violazioni delle leggi sui titoli. Questa notizia ha contribuito a un aumento dell'incertezza e del rischio percepito dagli investitori riguardo alla stabilità e alla conformità legale di FTX [21].
- 19 ottobre 2022: Sam Bankman-Fried ha annunciato che FTX aveva intenzione di espandere i propri investimenti politici, donando significative somme di denaro a candidati e cause politiche negli Stati Uniti. Questo ha sollevato interrogativi su possibili conflitti di interesse e ha accentuato il controllo regolamentare su FTX [22].

Questi eventi hanno portato a un aumento dei commenti sui post riguardanti FTX, con picchi significativi di attività che riflettono le preoccupazioni crescenti degli investitori. La media dei commenti ritorna a valori simili a prima delle due notizie, ma la deviazione standard rimane più alta, indicando che l'argomento è diventato più discusso e controverso.

Poco prima del crollo, in concomitanza con l'annuncio del CEO di Binance sulla vendita totale di FTT, si osserva un picco sia nella media che nella deviazione standard dei commenti. Dopo il crollo, la media si abbassa, rimanendo comunque più alta rispetto ai mesi precedenti, mentre la deviazione standard diminuisce leggermente ma rimane elevata per un lungo periodo.

4.6 Modellazione e Previsione

Un modello di regressione random forest [28] è stato creato e addestrato utilizzando scikit-learn [7]. Le feature sono state normalizzate e il modello è stato valutato utilizzando la Mean Squared Error (MSE) [27].

4.6.1 Preparazione dei Dati

I dati sono stati raccolti e pre-elaborati, includendo post e commenti (numeri e sentiment) sui social media e dati di Google Trends relativi agli eventi di interesse (crollo di Terra e FTX). Questi dati sono stati organizzati in un unico DataFrame e normalizzati su una scala da 0 a 1 per garantire che tutte le feature avessero lo stesso peso durante l'addestramento del modello.

I dati sono stati salvati in due file CSV, uno per il crollo di Terra e uno per il crollo di FTX, contenenti le stesse feature.

date	mean_comme	var_comm	std_comme	upper_std	lower_std	comm_sco	num_comm	score_mean	num_comn	num_post	negative	neutral	positive	trend	price_luna
01/04/2022	2,98203E+14	1,9758E+14	1,40563E+15	1,74633E+14	1,0115E+15	5,53852E+15	2,05724E+15	3,94145E+15	2,17922E+14	1,28859E+15	7,11377E+15	6,38254E+14	9,14815E+15	5,45455E+14	1030043686
02/04/2022	3,99519E+15	7,2586E+14	2,69417E+15	3,09565E+14	2,22973E+15	6,71558E+14	1,67953E+15	1,66834E+14	3,59281E+15	8,18792E+14	6,69067E+15	6,53702E+15	9,36114E+14	4,54545E+14	10587786293
03/04/2022	5,36904E+15	8,1657E+14	2,85758E+15	3,44884E+15	2,17358E+15	6,61591E+15	1,69747E+15	4,54469E+15	3,72211E+15	9,93289E+14	6,86266E+15	6,42631E+15	9,29386E+15	5,45455E+14	11517942093
04/04/2022	4,54461E+15	5,6037E+14	2,3672E+15	2,87117E+15	1,7842E+14	5,01276E+15	2,20797E+15	2,87713E+15	3,17677E+15	1,20805E+15	6,71937E+15	6,24315E+15	9,46088E+14	4,54545E+14	11288947256
05/04/2022	3,67667E+15	5,044E+14	2,24589E+15	2,63119E+15	1,80017E+15	5,27081E+15	1,97111E+15	1,51809E+15	2,88487E+15	1,58389E+15	5,9864E+15	6,33231E+15	9,90177E+14	5,45455E+14	11640347464
06/04/2022	4,152E+15	9,7181E+14	3,11739E+15	3,51311E+15	2,65961E+14	5,78491E+15	2,17926E+15	5,77141E+15	3,29972E+15	1,54362E+15	6,85901E+15	6,13079E+14	9,41531E+15	9,09091E+14	11624874557
07/04/2022	4,86105E+14	8,1851E+14	2,86096E+14	3,37772E+15	2,26315E+15	7,33035E+15	2,26E+15	1,71724E+15	3,78972E+14	8,5906E+14	6,89411E+15	5,9971E+14	9,44634E+14	4,54545E+14	10804476659
08/04/2022	3,40945E+15	4,6103E+14	2,14716E+15	2,50008E+15	1,73889E+15	5,28349E+15	1,47856E+14	3,58059E+15	3,03541E+15	1,20805E+15	7,01912E+15	7,11363E+15	8,91515E+15	5,45455E+14	10352425537
09/04/2022	4,9841E+15	6,9421E+13	2,63479E+15	3,18488E+15	1,99844E+15	5,39793E+15	1,39063E+14	3,87131E+15	5,1745E+15	8,05369E+14	5,97816E+15	6,86571E+15	9,69221E+15	4,54545E+14	9455789448
10/04/2022	3,70203E+15	4,4326E+13	2,10538E+14	2,50391E+15	1,64434E+15	5,05852E+15	1,67235E+15	1,94667E+15	3,07783E+15	1,04698E+15	5,9926E+15	6,07849E+15	10	3,63636E+14	9742293621
11/04/2022	6,71615E+15	1,6574E+15	4,07115E+15	4,7771E+15	3,25448E+15	5,43532E+15	2,007E+15	8,10229E+15	5,80838E+15	1,19463E+15	7,4704E+15	6,32126E+15	8,94084E+15	3,63636E+14	9186083808
12/04/2022	4,23375E+15	9,0987E+14	3,01641E+15	3,43093E+15	2,53689E+15	4,632955E+15	2,14696E+15	2,5675E+15	3,2905E+15	1,2349E+14	6,96076E+15	6,08736E+15	9,36662E+14	3,63636E+14	8249235132
13/04/2022	3,88781E+14	4,9474E+14	2,22428E+15	2,64191E+15	1,74115E+15	7,6985E+14	1,88229E+15	3,9973E+15	3,30331E+15	1,32886E+15	7,24652E+15	6,33884E+15	9,07939E+15	5,45455E+14	8456653928
14/04/2022	4,54809E+15	7,4577E+14	2,73088E+15	3,2107E+15	2,1758E+15	5,06252E+15	1,75758E+14	5,54696E+15	4,01962E+15	8,99329E+14	6,24536E+15	6,86855E+15	9,51725E+14	5,45455E+14	8791405389
15/04/2022	5,71758E+15	6,8126E+13	2,61009E+15	3,26908E+15	1,84774E+15	6,4194E+15	1,95765E+15	4,59746E+15	4,61183E+15	9,12752E+14	6,51514E+15	6,69869E+13	9,41018E+15	5,45455E+14	8155973691
16/04/2022	4,06448E+15	3,8662E+14	1,96627E+15	2,42722E+15	1,43303E+15	4,98404E+15	1,53957E+15	1,43307E+15	3,20457E+15	7,91946E+14	6,54408E+15	6,2627E+15	9,56703E+15	5,45455E+14	8034948902
17/04/2022	4,7268E+15	3,7828E+13	1,94493E+15	2,50416E+15	1,298E+15	8,03592E+15	1,56917E+15	6,48864E+15	3,87547E+15	1,03356E+15	6,61237E+14	6,23318E+15	9,53449E+15	6,36364E+14	8094802061
18/04/2022	6,61206E+15	1,2221E+14	3,49587E+15	4,2256E+15	2,6517E+15	6,2668E+15	2,50583E+15	5,45547E+15	4,65984E+15	1,30201E+15	6,92658E+15	6,6386E+15	9,16675E+15	5,45455E+14	7720529402
19/04/2022	4,02167E+15	5,1824E+13	2,27648E+15	2,71014E+15	1,7748E+15	5,26907E+15	1,74412E+14	2,65297E+15	3,0043E+15	1,2349E+14	6,67678E+15	6,08873E+15	9,5508E+15	4,54545E+14	9107966002
20/04/2022	5,33953E+15	6,6491E+14	2,57858E+15	3,18444E+15	1,8777E+15	5,6068E+15	1,75668E+15	6,15764E+14	4,30075E+15	1,03356E+15	7,1764E+14	6,17406E+15	9,19141E+15	3,63636E+14	9547008023

Figura 16: Parte iniziale del csv di Terra

4.6.2 Normalizzazione delle Feature

La normalizzazione delle feature è stata effettuata utilizzando il metodo MinMaxScaler di scikit-learn. Questo metodo trasforma i valori delle feature in modo che rientrino nell'intervallo [0, 1], permettendo di mantenere la proporzione tra i vari valori e facilitando il processo di addestramento del modello.

4.6.3 Random Forest Regression

Per la modellazione è stato utilizzato il RandomForestRegressor di scikit-learn. Il processo eseguito per ogni caso comprende diverse fasi comuni:

- **Preparazione delle Feature e del Target**: I dati sono stati divisi in feature (X) e target (y). Le feature includevano variabili come il volume degli scambi, i punteggi del sentiment, e altre metriche rilevanti, mentre il target era rappresentato dal prezzo della criptovaluta.
- **Addestramento del Modello**: Il modello RandomForestRegressor è stato addestrato sui dati di training. Questo modello utilizza un insieme di alberi decisionali per effettuare predizioni accurate e ridurre il rischio di overfitting.
- **Valutazione del Modello**: Le performance del modello sono state valutate utilizzando la Mean Squared Error (MSE) e il coefficiente di determinazione R^2 sul set di test. Queste metriche forniscono una misura della precisione delle predizioni del modello.

4.6.4 Training e test sullo stesso dataset

I dati sono stati suddivisi in un set di addestramento e un set di test utilizzando il metodo train_test_split, con una percentuale di test del 20% e un random state di 42. Questo consente di valutare le performance del modello su dati non visti durante l'addestramento.

I grafici nelle Figure 17 e 18 mostrano le previsioni ottenute dal modello addestrato e testato sui dati del crollo di Terra e FTX rispettivamente.

MSE su Terra: 106.94254489664006
R^2 **su Terra**: 0.9435246999507483

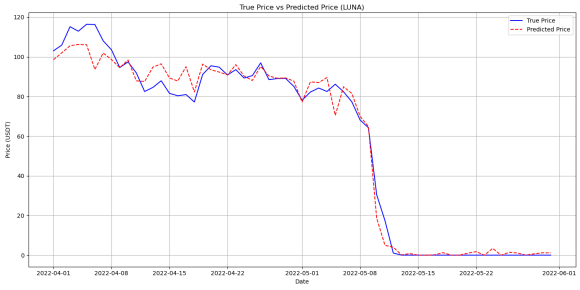

Figura 17: Previsione su LUNA

MSE su FTX: 15.55796913863171
R^2 **su FTX**: 0.8944551334809964

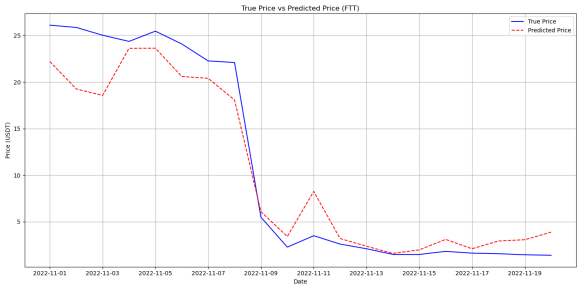

Figura 18: Previsione su FTT

4.6.5 Training e Test su Dataset Separati

Il modello è stato addestrato su un intero dataset (ad esempio, FTX) e poi testato su un altro dataset completo (ad esempio, Terra). Questo metodo consente di valutare la capacità del modello di generalizzare su dati completamente nuovi e differenti.

- Definizione di una griglia di parametri per ottimizzare il modello Random Forest utilizzando Grid Search [29].
- Addestramento del modello sui dati normalizzati del dataset FTX e poi test sui dati normalizzati del dataset Terra (e viceversa).
- Denormalizzazione delle previsioni per confrontarle con i prezzi reali e valutazione del modello utilizzando la MSE e il R^2.

MSE su Terra: 1524.4159749599964
R^2 **su Terra**: 0.206616142503391

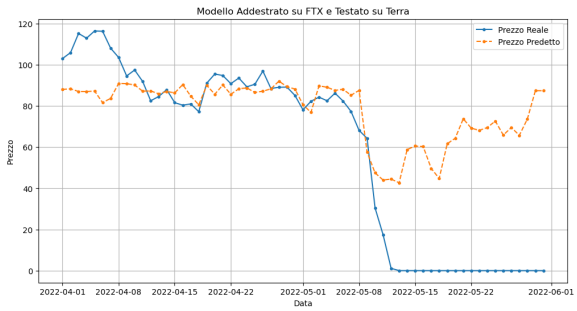

Figura 19: Modello addestrato su FTX e testato su Terra

MSE su FTX: 30.454951292439688
R^2 **su FTX**: 0.745211567403987

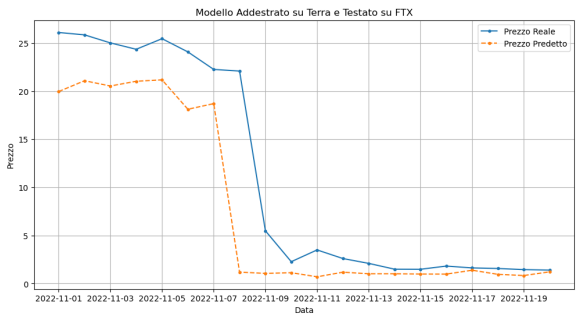

Figura 20: Modello addestrato su Terra e testato su FTX

4.6.6 Training Iterativo

Per ogni giorno, il modello è stato addestrato utilizzando tutti i dati disponibili fino a quel giorno. La previsione è stata poi effettuata per il giorno successivo. Questo processo è stato ripetuto iterativamente per tutto il dataset, partendo da un periodo iniziale di alcuni giorni di dati storici.

- È stata impostata una dimensione iniziale di alcuni giorni per il training set (10 per Terra e 4 per FTX).
- Per ogni giorno successivo, il modello è stato addestrato con i dati disponibili fino a quel giorno e il prezzo del giorno successivo è stato previsto.
- Le previsioni sono state denormalizzate per confrontarle con i prezzi reali e il modello è stato valutato utilizzando la MSE e il R^2.

MSE su Terra: 787.3589854901816
R^2 **su Terra**: 0.5458491325589196

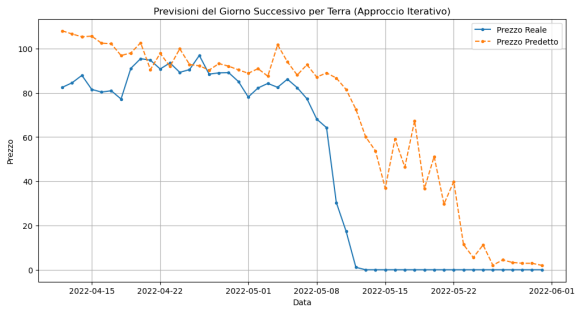

Figura 21: Previsioni del giorno successivo per Terra (approccio iterativo)

MSE su FTX: 103.85312402549131
R^2 **su FTX**: -0.5063670563093874

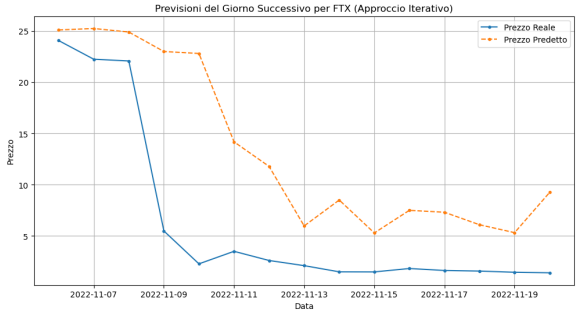

Figura 22: Previsioni del giorno successivo per FTX (approccio iterativo)

4.6.7 Conclusioni

L'uso della Random Forest Regression per la predizione dei prezzi delle criptovalute ha mostrato risultati promettenti, con una buona capacità di adattamento ai dati di addestramento e una ragionevole accuratezza nelle previsioni. Le previsioni incrociate, sebbene meno precise, indicano che il modello può generalizzare parzialmente a dati non visti.

Le previsioni future, tuttavia, hanno prodotto risultati meno soddisfacenti, indicando che c'è ancora molto lavoro da fare prima che possano essere utilizzate efficacemente in applicazioni pratiche per la previsione di crisi di mercato.

4.7 Immagini riassuntive

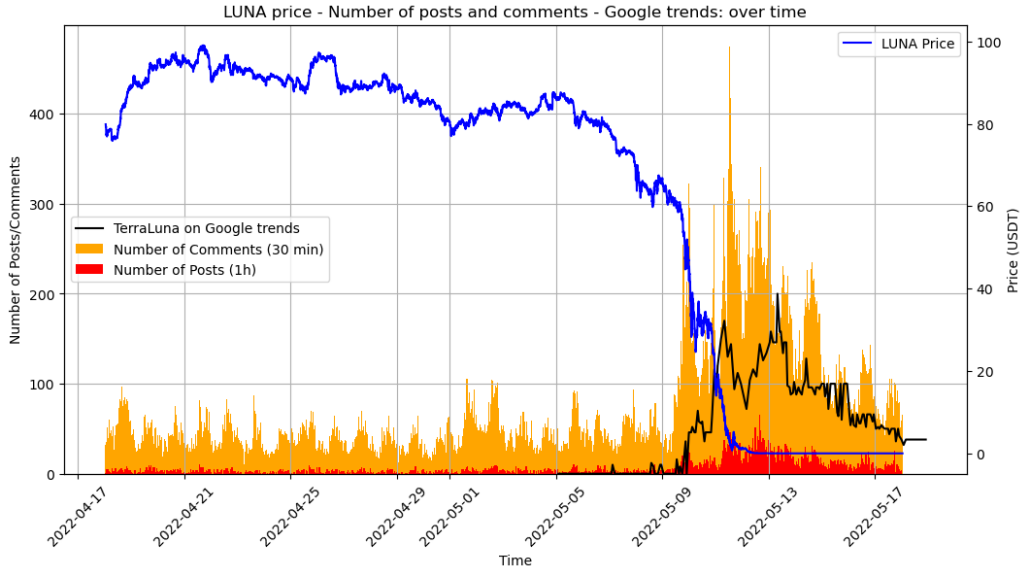

Figura 23: Terra

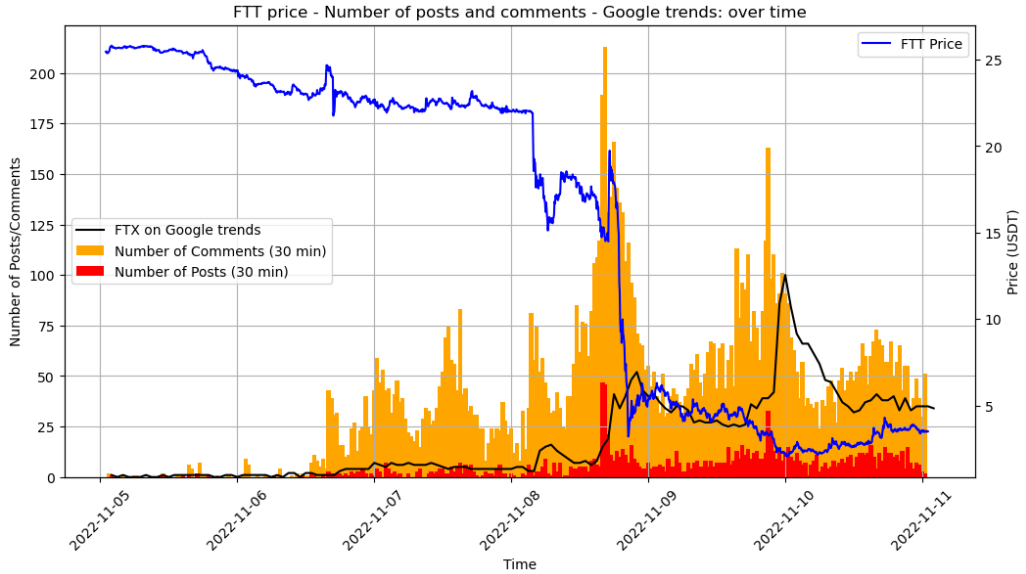

Figura 24: FTX

5 Conclusioni

In questo studio, abbiamo esplorato l'utilizzo dei dati dei social media, in particolare quelli raccolti dal subreddit `CryptoCurrency`, per analizzare il sentiment del mercato delle criptovalute e prevedere i potenziali crolli. L'obiettivo era di fornire uno strumento aggiuntivo per gli investitori, sia trader attivi che "holder", per identificare segnali di allarme tempestivi.

I risultati ottenuti dalle varie analisi indicano che:

- **Analisi del Sentiment**: Utilizzando VADER per l'analisi del sentiment, abbiamo riscontrato che un aumento dei commenti negativi può essere un indicatore di un imminente crollo del prezzo delle criptovalute. Tuttavia, è emerso che l'analisi del sentiment è meno utile quando i dati sono scarsi, come nel caso di FTX prima del crollo, rispetto a quando c'è già una significativa discussione sull'argomento, come nel caso di Terra.

- **Integrazione con Google Trends**: L'analisi delle ricerche su Google Trends ha mostrato una correlazione tra l'aumento delle ricerche e i periodi di crisi delle criptovalute. Questo suggerisce che l'interesse di ricerca su Google può essere un indicatore tempestivo delle crisi finanziarie nelle criptovalute, fornendo un ulteriore livello di analisi rispetto ai soli dati dei social media.

- **Andamento dei Commenti per Post**: Il calcolo della media e della deviazione standard giornaliera dei commenti per post ha rivelato pattern significativi di attività. Nei periodi di crollo, come evidenziato dai casi di FTX e Terra, si osserva un aumento della variabilità nel numero di commenti, riflettendo l'aumento delle discussioni e delle preoccupazioni tra gli investitori.

- **Modellazione e Previsione**: L'uso della Random Forest Regression per la predizione dei prezzi delle criptovalute ha mostrato risultati promettenti, con una buona capacità di adattamento ai dati di addestramento e una ragionevole accuratezza nelle previsioni. Le previsioni incrociate, sebbene meno precise, indicano che il modello può generalizzare parzialmente a dati non visti. Le previsioni future, tuttavia, hanno prodotto risultati meno soddisfacenti, indicando che c'è ancora molto lavoro da fare prima che possano essere utilizzate efficacemente in applicazioni pratiche per la previsione di crisi di mercato.

In conclusione, l'analisi dei dati dei social media combinata con strumenti di machine learning e dati di Google Trends offre un approccio innovativo e potenzialmente efficace per prevedere i movimenti di mercato delle criptovalute. I risultati sono promettenti, tuttavia ulteriori ricerche e miglioramenti del modello sono necessari per aumentarne l'affidabilità e l'accuratezza.

Questo studio pone le basi per future esplorazioni in questo campo, con l'obiettivo di fornire strumenti sempre più sofisticati e utili per gli investitori nel mercato delle criptovalute. Tra i lavori futuri, potrebbe essere utile effettuare l'analisi predittiva con una maggiore granularità temporale, ad esempio su base oraria, per cogliere variazioni più rapide e dettagliate. Inoltre, sarebbe interessante estendere l'analisi ad altri casi studio per valutare la generalizzabilità del modello e migliorare la robustezza delle previsioni.

Riferimenti bibliografici

[1] Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting, MDPI. Available: https://www.mdpi.com/2504-2289/7/3/137

[2] Cryptocurrency Price Prediction using Twitter Sentiment Analysis, arXiv. Available: https://arxiv.org/pdf/2303.09397

[3] Forecasting Cryptocurrency Returns from Sentiment Signals: An Analysis of BERT Classifiers and Weak Supervision, arXiv. Available: https://arxiv.org/pdf/2204.05781

[4] A Comprehensive Study on Cryptocurrency Price Prediction Using Machine Learning and Social Sentiment, Springer. Available: https://link.springer.com/article/10.1007/s10489-022-03241-9

[5] Pandas Documentation. Available: https://pandas.pydata.org/

[6] Matplotlib Documentation. Available: https://matplotlib.org/

[7] scikit-learn Documentation. Available: https://scikit-learn.org/

[8] Getting Reddit Data for Academic Research, Pushshift. Available: https://www.reddit.com/r/pushshift/comments/1b331ho/getting_reddit_data_for_academic_research

[9] Academic Torrents. Available: https://academictorrents.com/details/56aa49f9653ba545f48df2e33679f014d2829c10

[10] r/CryptoCurrency, Reddit. Available: https://www.reddit.com/r/CryptoCurrency/

[11] Hutto, C.J. and Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available: https://github.com/cjhutto/vaderSentiment

[12] Google Trends. Available: https://trends.google.com/

[13] Binance to Sell Rest of FTX Token Holdings as Alameda CEO Defends Firm's Financial Condition, CoinDesk. Available: https://www.coindesk.com/business/2022/11/06/binance-to-sell-rest-of-ftx-token-holdings-as-alameda-ceo-defends-firms-financial-co

[14] Binance to liquidate all FTX Tokens as Alameda CEO defends financials, Cointelegraph. Available: https://cointelegraph.com/news/binance-to-liquidate-all-ftx-tokens-as-alameda-ceo-defends-financials

[15] FTX assures that 'assets are fine' and contends withdrawal complaints, Cointelegraph. Available: https://cointelegraph.com/news/ftx-assures-that-assets-are-fine-and-contends-withdrawal-complaints

[16] Sam Bankman-Fried Says 'FTX Is Fine.' Investors Say, 'I Doubt It.', CoinDesk. Available: https://www.coindesk.com/business/2022/11/07/sam-bankman-fried-says-ftx-is-fine-investors-say-i-doubt-it/

[17] Binance. Prezzi delle Criptovalute. Disponibile: https://www.binance.com/it/landing/data

[18] CoinDesk. Terra's Luna Cryptocurrency Falls to Nearly $0 as UST Stablecoin Collapses. Available: https://www.coindesk.com

[19] CoinCodex. Why Did Luna Crash 99.99%? Here's What Happened to Luna. Available: https://www.coincodex.com

[20] Richmond Fed. The Collapse of Terra: The Looming Stablecoin Crisis. Available:https://www.richmondfed.org

[21] SEC Investigates FTX for Potential Violations of Securities Laws. Available: https://www.sec.gov/news/press-release/2022-201

[22] The Week. Sam Bankman-Fried's Political Donations Raise Questions. Available: https://www.theweek.com/articles/976043/sam-bankman-fried-political-donations-raise-questions

[23] Curve.fi è un protocollo di finanza decentralizzata. Available: https://curve.fi/

[24] Perdere il peg di una valuta significa che la valuta non mantiene più il suo valore fisso rispetto a un'altra valuta.

[25] UST (TerraUSD) era una stablecoin algoritmica legata al valore del dollaro statunitense, parte dell'ecosistema Terra, progettata per mantenere la parità con il dollaro attraverso un sistema di arbitraggio con la criptovaluta LUNA.

[26] LFG (Luna Foundation Guard) era un'organizzazione non profit creata per sostenere e stabilizzare l'ecosistema Terra, in particolare la stablecoin UST, attraverso riserve di criptovalute.

[27] MSE: Fornisce una misura dell'errore medio quadratico tra i valori predetti e quelli effettivi. Un valore più basso indica una maggiore precisione delle predizioni.

R^2: Indica quanto bene il modello spiega la variabilità dei dati osservati. Un valore più alto (più vicino a 1) indica un modello che spiega meglio la variabilità dei dati.

[28] La Random Forest è un modello di machine learning basato su un insieme di alberi decisionali, utilizzato per compiti di classificazione e regressione, che migliora la precisione e riduce il rischio di overfitting combinando le predizioni di molti alberi diversi.

[29] Grid Search è una tecnica di ottimizzazione degli iperparametri utilizzata in machine learning. Consiste nel provare sistematicamente tutte le combinazioni di un insieme predefinito di parametri per trovare la configurazione che offre le migliori prestazioni del modello.